Representation Learning for Continual Task Performance

Timo Flesch

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University of Oxford

New College

University of Oxford

Trinity Term, 2022

Abstract

Humans have the remarkable ability to learn continually without forgetting, and adapt their behaviour to changing situational demands. While previous work has focussed on elucidating the mechanisms that underlie flexible context-dependent processing of information, much less is known about the format in which information is represented in the human brain, and how this promotes continual task performance. The aim of this DPhil was to develop computationally informed theories of representation learning for context-dependent processing, and test these in behavioural and neuroimaging recordings from healthy human participants.

Across a series of neural network simulations, behavioural and neuroimaging studies and re-analyses of freely available datasets with recordings from macaque FEF, I gathered evidence in support of earlier theories of cognitive control, which postulated that prefrontal cortex implements gating strategies that favour task-relevant over task-irrelevant information in the service of context-specific task goals. In chapter 3, I propose a computational framework to study representation learning for context-dependent decisions with artificial neural networks, and demonstrate how the same architecture can learn either high-dimensional and task-agnostic or low-dimensional and task-specific representations. In chapter 4, I tested predictions from these simulations in fMRI recordings from human participants who learned to perform a similar context-dependent decision task, finding that representations in fronto-parietal regions were highly task-specific, with relevant information from distinct tasks mapped onto orthogonal coding axes. In chapter 5, I introduce a model of human continual learning, in which the gating signal is learned by a simple Hebbian mechanism. Lastly, in chapter 6, I tested whether previously reported benefits of blocked over interleaved training

generalised to abstract rules and whether they promote cross-domain transfer.

Taken together, this thesis introduces a computational theory of continual representation learning and provides empirical evidence that the human brain uses gating strategies to represent relevant information in context-specific subspaces.

Statement of Contributions

The thesis begins with a general introduction and a review of the relevant literature (Chapter 1 & 2), and ends with a general conclusion. These chapters are independent work. In the following, I describe how others contributed to the work laid out in the empirical chapters of this thesis.

Research presented in all empirical chapters (Chapters 3-6) was primarily conducted by me, under the supervision of Prof. Christopher Summerfield, who provided input on the conceptualisation, design, analysis and writing.

A condensed version of Chapter 3 and 4 has been published in the journal Neuron, with myself as first author (https://doi.org/10.1016/j.neuron.2022.01.005). I wrote the initial draft of the manuscript and revised it together with my supervisors Prof. Summerfield and Dr. Andrew Saxe.

Chapter 3: The distinction between rich and lazy learning and the gating theory were developed under close conceptual guidance by my second supervisor, Dr. Saxe. All simulations, analyses and figures were implemented and prepared by me.

Chapter 4: The fMRI study was conceptualised and designed together with my supervisor Prof. Summerfield. I programmed the experiment, collected and analysed the data and prepared the figures and manuscript. Dr. Keno Juechems and Dr. Tsvetomira Dumbalska translated all task instructions of the fMRI study to Spanish assisted with the participant recruitment and collection of the fMRI data in Granada. I obtained

additional help from the late Tania Martínez Montero during data collection. The NHP data was analysed by me. Dr. Andrew Saxe developed the RNN model described in Appendix B.

Chapter 5: This work was initially conducted in collaboration with a visiting PhD student, David G. Nagy, who was supervised by me and Prof. Christopher Summerfield. I conceptualised the model together with Prof. Summerfield. An initial version of the model was then implemented by David G. Nagy under my supervision, but further simplified and refined by myself. All simulations, analyses and figures presented in the chapter were implemented and prepared by me. A manuscript based on the chapter is currently under review and has been published as preprint on Arxiv, with co-first authorship shared between me and David G. Nagy (https://doi.org/10.48550/arXiv.2203.11560). I wrote the paper together with my supervisor and incorporated comments from David G. Nagy and Dr. Saxe.

Chapter 6: Prof. Summerfield, Dr. Paul Muhle-Karbe and I conceptualised and designed the experiments. Dr. Paul Muhle-Karbe assisted with the collection and initial norming of the stimulus sets. I programmed all experiments and collected and analysed the data. The neural network model described in Appendix C was developed by Cal Shearer as part of her MSc project which she carried out under supervision by Prof. Summerfield and myself.

Acknowledgements

I am very grateful to my primary supervisor Chris Summerfield for being a supportive mentor who challenged my thinking and provided the intellectual nurturing for me to grow from an ambitious, but rather clueless undergraduate student, to a slightly less clueless DPhil candidate. Thank you for teaching me perseverance when results were puzzling and unexpected, for encouraging me to try my best and run yet another exciting analysis in the face of looming deadlines, for introducing me to numerous other scientists around the globe and enabling me to present my findings at several international conferences, for fun lab retreats and even more entertaining Chris(t)mas parties. I would also like to express my gratitude to Andrew, my second supervisor, for introducing me to the world of deep learning theory, and for teaching me the beauty of small, tractable models, which had a fundamental impact on my thinking. I am incredibly lucky to have been able to work under the guidance of such an exceptional supervisory team.

When someone asks me for how many years I've been living in Oxford, I usually say that I joined the Summerfield lab long before Chris Summerfield himself. Having been able to spend so many years in the same fantastic lab - first as visiting student, then as RA and finally as DPhil - is a privilege I'm very grateful for. Due to the rather short half-life of the average non-tenured academic, many great minds who contributed to my work in one way or the other have long left. I'm grateful to every single member of the Summerfield lab, past and present, but would like to use this limited space to mention a few of you directly. Keno, thanks for being a great friend and for all the support over the years! I would like to thank Hamed for introducing me to the magic

world of RSA, and for an endless supply of jokes of the highest quality. Thanks to Jan for being my unofficial third supervisor during my days as little undergraduate student and your contagious drive and enthusiasm. Paul, thank you for being a fantastic collaborator on a challenging project we're both passionate about, and for advice on parenting. Fabrice, "our paths have run parallel for most of our time in Oxford". You've been a great lab mate, and I still have fond memories of our jam sessions in the New College band room. Ron, thanks for sharing my passion for trees. Mira, thanks for a great trip to (the fMRI scanning facilities in) Granada and for deeply caring about the world, even if this meant that you'd have to drag me (with help from Keno) to the same vegan restaurant every single day. The crisps weren't too bad.

I would also like to thank my friends in Germany, in particular my buddies in the wild west, who somehow managed to stay in touch. Thanks to my family, in particular my mum Lydia, for encouraging and supporting me to follow my dreams, even when they took me to another country. This thesis would not have been possible without your support. Leonie, thanks for growing fond of me despite the initial shock that I'm also from Germany, and for making every single day special. I came to Oxford for the science, but found something much more valuable. Lastly, thanks to Arlo, for putting many things into perspective and showing me what's really important in life.

This work was made possible by a Medical Sciences Graduate School Studentship, jointly funded by the Medical Research Council and the Department of Experimental Psychology. I received further support from the COVID-19 Scholarship Extension Fund, offered by the University of Oxford. Indirect support was provided through grants awarded to Prof. Summerfield. I am grateful to the funding agencies for their support.

Contents

List of Figures

General Introduction

Throughout our lifetime, we continuously expand our understanding of the world around us. For example, we might first learn how to judge fruit by its shape, enabling us to distinguish between different types of fruit, and later by its colour, to decide whether some fruit is ripe. Acquiring such a sophisticated understanding of various objects enables us to respond flexibly to ever evolving situational demands. For instance, on a trip to a supermarket, our goal might be to buy a certain type of fruit, and we might care less about whether the fruit is already ripe. A few days later, at the breakfast table, when we're looking for some fruit to accompany our healthy meal, we might explicitly want to choose ripe fruits. This requires us to judge the same type of stimulus according to different criteria, focussing on different perceptual attributes (and/or actions) in different contexts. The aim of this thesis is to study the computations and representations that support this continual task performance.

How does the brain coordinate this sophisticated context-appropriate behaviour? One possibility would be that it has learned stimulus-response associations which determine an appropriate response for every conceivable stimulus. Human behaviour, however, seems to be much more sophisticated than that, as we are able to override habitual responses in the service of context-specific task goals, suggesting that we can actively control the way in which we interact with the world (Posner & Snyder, 1975). This realisation that *thinking* plays a fundamental role in our interactions with the world was reflected in a paradigm shift from Behaviourism, which dominated Psychology in the first half of the last century and focussed on habitual stimulus-response associations (Pavlov, 1927; Skinner, 1950; Thorndike, 1932; Watson, 1913), to Cognitivism (Miller et al., 1960; Newell et al., 1958). Central to this idea is the assumption

that cognitive processes can be understood in terms of representations and computations that are carried out on those representations (Thagard, 2020). In other words, Cognitivists proposed that thinking can be formalised as a type of information processing, a doctrine which was heavily influenced by the cybernetic movement of the 1950s and the then growing interest in Artificial Intelligence (McCulloch & Pitts, 1943; Minsky, 1961).

Over the following decades, Psychologists became increasingly invested in the study of the control processes that govern flexible perception and action selection, leading to the notion of *cognitive control*, an umbrella term for the brain's ability to exert top-down influence on the way in which we perceive and interact with our surroundings (Posner & Snyder, 1975). A substantial body of research has since investigated behavioural markers of cognitive control (Monsell, 2003) and possible neural correlates of its constituent processes such as error monitoring and resource allocation (Miller, 2000). However, much less is known about the relationship between the nature of *neural representations*, in other words the *geometry* of neural firing patterns, and task performance, and how these patterns are sculpted continually by experience, which is the central theme of this thesis. I will specifically focus on so-called context-dependent decision-making tasks, in which the same stimulus is judged according to criteria that depend on the context in which it occurs. The theme of this thesis can be divided into a set of sub-topics, each of which will be addressed in a separate chapter:

1. A computational model of context-dependent processing.

Following a (neo-)connectionist tradition (Rumelhart & McClelland, 1987; Richards et al., 2019), I will use deep artificial neural networks as models of representation learning and begin with a fundamental question: What are the relative costs and benefits of different representational schemes for context-dependent processing and under which circumstances should we expect to observe them? In **Chapter 3**, using simulations with deep artificial neural networks trained on context-dependent decision-making problems, I will explore trade-offs associated with representations that are either tailored to specific task demands (task-specific) or much more general

(task-agnostic), and analyse how they might emerge.

2. Neural correlates of context-dependent processing.

In **Chapter 4**, I present results from a study in which we investigated the geometry of task representations in the brain. I will describe a neuroimaging study with healthy human participants and a supplementary re-analysis of a dataset with single- and multi-unit recordings from the macaque brain which provide converging evidence that the brain might use task-specific and low-dimensional representations to perform optimally in different contexts.

3. A model of human continual learning.

A key limitation of the modelling work presented in Chapter 4 is that standard neural networks can't learn continually without forgetting how to perform previously learned tasks. A simple solution which I also adapted for my simulations was to train the network on randomly interleaved data. However, the literature on cognitive control has demonstrated that rapid switches between tasks are associated with a cost, which suggests that the brain has evolved mechanisms to learn particularly well when information is presented continually and switches between tasks occur rarely. In other words, time itself appears to serve as a strong cue to guide the formation of hierarchical context-dependent representations. In **Chapter 5**, I present a simple model of human continual learning, which can learn two tasks sequentially without interference and performs worse when training data is randomly interleaved.

4. Abstract representations for cross-domain transfer.

Most of the work discussed in this thesis is concerned with perceptual decisions. However, many decisions in the real world are not directly grounded in the visual appearance of objects, such as the dangerousness of an animal, or the maximum speed of a vehicle. These concepts are often truly abstract, in the sense that they can be applied to objects from many different domains. We can judge whether a vehicle is fast or slow, in the same way in which we can ask whether an animal is fast or slow. In

Chapter 6, I explore in a series of behavioural experiments whether findings presented in previous chapters generalise to more abstract concepts. More specifically, I tested whether blocked (or continual) learning, in contrast to interleaved learning, facilitates the formation of abstract representations that are generalisable across domains.

In the following two chapters, I provide a general review of extant literature on cognitive control and representation learning, and introduce neural networks as toolkit to model representation learning. In **Chapter 1**, I will begin with an overview of the literature on cognitive control and its neural correlates, covering early theoretical work and empirical investigations, both in behaviour and at the neural level. This is followed by a review of the literature on neural representations. Next, in **Chapter 2**, I will introduce deep neural networks as models of representation learning. I will begin with a brief primer on neural network architectures and their training methods, followed by an overview of studies that have discovered links between the way in which information is represented in deep neural networks and biological brains. I will end this literature review with a brief discussion of a key limitation of contemporary neural network models, with a particular focus on the inability to learn continuously without catastrophic forgetting. The chapter shall conclude with a statement of the main objectives of this thesis and briefly revisit the four topics outlined above in light of the discussed literature.

Chapter 1

Cognitive control and the prefrontal cortex

As we navigate our everyday life, the brain has to process a seemingly endless stream of continually evolving and context-specific information. Yet, we can selectively prioritise relevant perceptual information, memories, or actions in response to these ever-changing situational demands. How does the brain coordinate this sophisticated context-appropriate behaviour? Theorists have argued that this requires top-down signals that determine how information is manipulated to shape behaviour, an idea that is referred to as executive functions or cognitive control (Posner & Snyder, 1975) and is thought to rely on the integrity of the prefrontal cortex (Miller, 2000). In this chapter, I will begin with a historical overview of influential theories on cognitive control, followed by evidence from neuroscience in support of these theories. I will then turn to the notion of *neural representations* and ask how task-relevant information is represented in the brain.

1.1 Cognitive control

Why is cognitive control needed at all? A classical finding in Psychology which illustrates the trade-off between automaticity (the execution of habitual responses) and the need of control is the so-called Stroop effect: Here, participants see written words that describe different colours. The colour of the ink either matches the word (e.g. the word "red" shown in red) or not ("red" shown in blue). Performance is worse on conflict tri-

als, for example when the word ”red” is displayed in blue and participants are asked to name the colour, in which case they tend to read out the word (Stroop, 1935). In other words, cognitive control is needed whenever situational demands require a response that deviates from a well-established mapping between inputs and outputs (Shiffrin & Schneider, 1977). In the next two sections, I explore theories of how these processes might be implemented in the brain, and evidence from neuroscience pointing at the crucial role of prefrontal cortex for continual task performance.

1.1.1 Theories of cognitive control

How can the brain cope with the plethora of information that it is exposed to every day? One solution would be to employ some form of *selective attention*, which would enable us to prioritise some percepts over others in the service of task-specific goals (Posner & Presti, 1987). Attention might require two different types of information processing pathways, one automatic pathway for well-established tasks, and another pathway that requires higher levels of cognitive control (Shiffrin & Schneider, 1977). A model which made this distinction explicit was proposed in the 1980s by Norman and Shallice. The authors assumed that in order to perform a specific task, humans choose from a set of “schemas” that determine information processing and action selection for a variety of different scenarios (Norman & Shallice, 1986). In their model, two processes control the formation and selection of these processing scripts, the contention scheduling system and the supervisory attention system (Norman & Shallice, 1986). The contention scheduling system controls the execution of automatic behaviour in response to well-known tasks. It is required to select an appropriate schema, which in turn ensures that task-relevant information is preferentially processed, and inappropriate actions are inhibited (Norman & Shallice, 1986). On top of this fast and automatic process acts the Supervisory Attention System (SAS), which is thought to implement cognitive control in less familiar situations or those with a particular need for response inhibition (Norman & Shallice, 1986). The authors postulate that the SAS may not only retrieve schemas from the “database” but might be crucial for the formation and implementation of novel schemas. While the Norman and Shallice model might be appealing for its simplicity, it is purely *symbolic* in the sense that it does not describe how

the brain might implement the two proposed mechanisms. It is exemplary of a tradition in psychology to describe mental processes as box-and-arrow diagrams, abstracting away from implementational details. A different approach is offered by so-called *connectionist* models of cognition, which formalise information processing with the help of artificial neural networks (Rumelhart & McClelland, 1987). Historically, neural networks were introduced as highly simplified models of the human brain, comprised of sets of units that are organised in layers and connected via weights. Those units and weights serve as high-level analogue to neurons and synaptic connections (Rumelhart & McClelland, 1987). In contrast to box-and-arrow diagrams, these artificial neural networks make a concrete proposal as to how cognitive function might be *implemented* in neural circuits and are exemplary of convergent research interests in Psychology and Artificial Intelligence (Rumelhart & McClelland, 1987). Probably the most influential Connectionist model of cognitive control was proposed by Miller & Cohen, (2001). They stipulated that the primary role of PFC is to exert top-down gain control on sensory and motor neurons, to selectively process task-relevant sensory information and prepare task-appropriate motor responses. The PFC achieves this modulation via b*bias signals* that control the flow of information across numerous cortical areas and thus facilitate task-appropriate mappings between inputs, internal representations, and outputs. But how can these biases be implemented? In Cohen et al., 1990, the authors proposed a simple neural network model of the Stroop task. It consisted of an input layer, encoding the sensory properties of the stimuli, a hidden layer, corresponding to a set of neurons with sigmoidal response profiles, and an output layer, representing the different actions that can be taken (such as naming a colour or reading a word). A fundamental property of sigmoid functions is that they saturate for very small and very large inputs, responding with zero or one respectively. In the model, Cohen et al. proposed that individual neurons might encode different sensory features (such as colour, or the meaning of a word). They introduced a separate task control layer which moves the activity of hidden units in and out of the saturating range, either maximally activating or wholly deactivating them. With this set-up, the role of cognitive control is to identify units that encode irrelevant/conflicting information (such as the colour

of a word during the naming task) and deactivate them via bias signals so that they no longer contribute to the response (Cohen et al., 1990). The model was heavily inspired by earlier work on visual attention (Treisman & Gelade, 1980) but subsequently developed further to generalise this notion of selective processing to encompass both sensory and motor signals, as well as memories and emotions (Miller & Cohen, 2001).

But how can the brain know whether a situation requires cognitive control? A potential answer to this question was provided by a model devised by Koechlin & Summerfield, (2007). Using concepts from information theory, they proposed to decompose the amount of information $H(a)$ required for selecting an action a into two terms, denoting the mutual information between a stimulus and action and the amount of information conditioned on the stimulus:

$$H(a) = I(s,a) + Q(a|s) \tag{1.1}$$

The equation describes the trade-off between automaticity (or what they refer to as sensorimotor control) and cognitive control in information-theoretic terms. If an action a is frequently preceded by a stimulus s, their respective mutual information would be high. In other words, the stimulus would convey most of the information required for triggering the action under consideration. The second term quantifies what is commonly referred to as cognitive control, the information required for selecting an action after accounting for sensorimotor control. It is low when the stimulus invariably triggers the action and high when the relationship between the two is weak. The authors went on to decompose $Q(a|s)$ further, into terms they refer to as contextual and episodic control:

$$Q(a|s) = I(c,a|s) + Q(a|s,c) \tag{1.2}$$

Contextual control determines how the task-context influences whether an action should be chosen, while episodic control measures the information conveyed by the recent (trial) history. Taken together, their model offers a mechanistic account of action selection, guided by a decomposition of executive function into different bottom-up

and top-down processing stages.

1.1.2 Neural correlates of cognitive control

How is cognitive control implemented in the brain? Early neuroscientific evidence suggests that the prefrontal cortex plays a pivotal role in the suppression of habitual responses and selection of task-appropriate actions. In this section, I will provide a short overview of early lesion experiments and more recent neuroimaging studies in support of this view.

Lesioning studies

The notion that prefrontal cortex is crucial for executive function can be traced back to Donald Hebb, who spent much of the 1930s studying the impact of temporal and prefrontal lobectomy on the cognitive ability of patients treated for epileptic seizures (Hebb, 1939). Interestingly, he first noticed that patients showed little performance deficits in the then-standard Binet test, which measured language-dependent abilities that showed only modest impairment (Hebb, 1939). In follow-up work, however, he reported that prefrontal lesions appeared to affect some cognitive abilities more than others. More specifically, in Hebb (1942), he proposed to divide intelligence into "present intellectual power" and established responses to routine problems. With this, he laid the foundation for what would later be introduced as intelligence A and B, or *fluid* and *crystalline* intelligence by Hebb and Catell respectively (Brown, 2016). Much of the subsequent work provided further evidence that lesions to prefrontal cortex affect raw cognitive power (Duncan, 1986; Luria et al., 1966). While basic abilities such as object recognition and motor control usually remain intact, patients tend to be unable to override habitual responses with context-appropriate behaviour or adapt to novel scenarios not experienced prior to the injury (Duncan, 1986; Luria, 1973), leading to error patterns described as "utilisation behaviour" and "perseveration behaviour". Utilisation behaviour marks the inability to suppress behavioural impulses in response to stimuli that are irrelevant to the achievement of the current goal (Lhermitte, 1983), such as when a patient habitually reaches for an object whenever it's in sight, irrespective of the task they were actually engaged in. Perseveration behaviour describes the tendency

to continue with a behavioural motif long after it has become irrelevant (Ridley, 1994). Since these early reports, numerous studies in controlled experimental settings have confirmed that task switch costs and deficits in inhibitory control are highly amplified in patients with prefrontal cortex damage (for a review, see Aron et al., 2004).

Neuroimaging

As lesions are the result of injury, their precise location as well as the extent of damage might vary considerably between patients (Rorden & Karnath, 2004). While some controlled lesion studies with non-human primates revealed comparable evidence (for example, Dias et al., 1997), the controlled investigation of prefrontal cortex function in human participants required different techniques. Functional magnetic resonance imaging (fMRI) offers the opportunity to investigate these processes in the healthy brain with decent spatial resolution. In brief, fMRI measures the difference in blood oxygenation between trials and rest, which is thought to be a proxy for the change in overall metabolic activity of neurons in response to changing task demands (Matthews & Jezzard, 2004). While there is an ongoing debate as to whether this *hemodynamic response* favours synaptic potentials or changes in axonal firing rates (Logothetis, 2003), it is well accepted in the community as reliable summary statistic of task-related neural activity (Poldrack, 2008).

Human fMRI studies offering further support for the role of PFC in cognitive control have reported that univariate activation of the PFC differs between task-stay and task-switch trials, i.e. those with the same or a different mapping between stimuli and responses (Liston et al., 2006; Woodward et al., 2006; Yeung et al., 2006). Interestingly, in these studies, neural correlates of task-switch effects weren't restricted to dorsal portions of PFC, with similar activity differences found in anterior cingulate and lateral parts of parietal cortex. As these regions seemed to be involved in a variety of different tasks relying on cognitive control, they are commonly referred to as multi-demand network (MD) in the literature (Duncan, 2010). Follow-up studies suggest functional subdivisions within this fronto-parietal network. In line with the theories discussed in the previous section, prefrontal cortex might implement a top-down modulation of task representations and preparation of motor responses, while the primary

role of ACC might be conflict monitoring during task performance (Botvinick et al., 2001; Kerns et al., 2004; MacDonald et al., 2000; Miller & Cohen, 2001).

But how is this top-down modulation implemented in PFC? Koechlin et al. (2003) devised a task that allowed them to investigate the relation between the information-theoretic quantities described in Koechlin & Summerfield (2007) and changes in the hemodynamic response, finding that different portions of PFC had functionally distinct roles. To remind the reader, their model divided executive function into sensorimotor, contextual, and episodic control processes, each of which was associated with a different information-theoretic quantity. In Koechlin et al. (2003) participants were asked to give responses to various coloured shapes and letters. The immediate context in which a stimulus appeared determined whether the stimulus and associated response remained constant or changed across blocks or from trial to trial. Using this design, they could for instance distinguish between different level of sensorimotor information, by showing a single stimulus and allowing only a single type of response per block ($I(s,a) = 0$) or several stimuli with matched number of response types ($I(s,a) = 1$). Neuroimaging revealed that the different information-theoretic quantities correlated with activity in distinct prefrontal regions, forming a processing cascade from anterior to posterior parts of PFC.

1.2 Task representations in the prefrontal cortex

The central tenet of cognitive science is that cognition can be understood in terms of representations and processes that operate on those representations (Thagard, 2020). While the previous sections have dealt with the *computations*, i.e. the control processes that govern task-appropriate behaviour, I will now focus on how task information is thought to be *represented* in the brain. Most of the early work on task representations followed the single-neuron doctrine, asking how individual neurons adapt their firing patterns to changing task demands (for a review see Miller, (2000)). The past decade, however, has been marked by a paradigm shift in neuroscience, driven by advances in recording techniques and statistical analysis tools, toward the so-called *population doctrine* which sees the distribution of activity in a group of neurons as the fundamental

unit of computation (Saxena & Cunningham, 2019; Yuste, 2015). Mathematically, a population of neurons can be described as multi-dimensional space, spanned by the firing rates of individual neurons. The ensemble of firing patterns that is elicited in response to a stimulus is then characterised by a position in this high-dimensional neural state space, which defines mathematically how information about the stimulus is *represented* in neural activity. The same concept can be applied to voxels from an fMRI study, electrode readings from an EEG study or units of an artificial neural network. In all of these cases, it is assumed that complex information is represented in the activity ensemble of a population of simpler units. This definition of neural representations lends itself to a few fundamental questions: What kind of information is encoded? And what is the geometry of this neural representation? In other words, where do different stimuli lie in the activity space? Are they orderly arranged, or do they appear in random positions? Is there a mapping between axes of variation in stimulus space (brightness, colour, shape, "dangerousness" etc.) and the arrangement of stimuli in the activity space? In this section, I will argue that while a substantial body of research has investigated what type of information is represented by prefrontal neurons, the research on the geometry of task representations is much more fragmented and inconclusive and requires further investigation.

1.2.1 The geometry of task representations in PFC

What information is encoded in task representations, and do these representations adapt to changing task demands? Most models of cognitive control reviewed in the previous section suggest that PFC biases internal representations in a task-dependent manner via modulatory signals. This implies that a task representation should encode information in a way that is conducive for task performance and prioritise relevant over irrelevant feature dimensions. An alternative would be that the neural code is highly heterogenous, with neurons coding for various idiosyncratic combinations of these feature dimensions. In this section, I briefly review evidence in favour of both of these views.

Task specific representations and axis-aligned neural codes

Early work with single-unit recordings from non-human primates (NHPs) provided evidence that representations are indeed highly task specific. Rainer et al. (1998) trained NHPs on a delayed-match-to sample (DMS) task involving arrays of three different objects of which only one was task-relevant on any given trial. Each trial began with three shapes displayed in different positions on the screen (sample phase). After a brief delay, another array with these shapes was displayed (test phase), with the shapes being either in the same (match trial) or different positions (non-match trial) on the screen. On each trial, only one of the shapes was relevant and monkeys were asked to remember it and perform a saccade to this shape in the test phase of the trial. Which shape was relevant was varied across blocks of trials and indicated by cue trials in which only a single shape was shown on the screen. The authors recorded spike rates from neurons in the lateral prefrontal cortex. While coding properties of these neurons were heterogenous, most units demonstrated highly task-specific responses, with activity depending on whether or not a particular shape was task-relevant. For instance, some neurons would only fire on trials in which a certain shape was relevant, and not respond to the other shapes or to that shape when it was irrelevant. Interestingly, this task-specific activity was already present before sample onset, indicating that the monkeys maintained the current task rule in memory.

Further evidence for task-specific neural activity was provided by the same group of authors in Asaad et al., (2000). Here, NHPs were trained to alternate between a spatial, object and associative task. In the spatial task, monkeys had to remember the location of a circle that was briefly shown either on the right or left side of the screen and perform a saccade to that remembered location after a short delay interval. In the object task, monkeys had to remember an image of an object shown at the beginning of a trial and were shown the same object and a new object after a short delay. Again, they were asked to perform a saccade to the remember object. Finally, in the associative task, monkeys had to learn an association between images of objects and either leftward or rightward saccades. Just like in the previous study, neural activity was recorded from lateral PFC. Overall, neural firing patterns were again heteroge-

nous, but the activity of many neurons was task-specific, with baseline firing rates and dynamics differing between the three tasks.

But what if different features of the same stimuli are task-relevant, depending on the context in which they occur? In Roy et al. (2010) NHPs where trained on a context-dependent variant of a delayed match to category task. Stimuli were images that varied in two dimensions, morphing them from cats to dogs along one dimension, and from one type of cat/dog to another along the other dimension. The authors applied two category boundaries to this bivariate stimulus space, a horizontal boundary in one context and a vertical boundary in another. Hence, in one context, monkeys had to decide whether the image depicted either a cat or a dog, whereas in the other, they had to distinguish between two different types of cats/dogs. Each trial began with a coloured dot that served as contextual cue, followed by two stimuli, the sample and probe, interspersed with a brief delay interval. The task was to decide whether the first and second stimulus belonged to the same context-specific category. This paradigm has the advantage of dissociating representations of stimuli from response-preparation signals, as the monkeys can categorise the first stimulus, but cannot prepare a response until they have seen the second stimulus. Behavioural findings suggested that the monkeys were able to learn the context-dependent nature of the task, with decisions predominantly being influenced by the relevant – but not the irrelevant – dimension in each context. The authors recorded single-cell activity from a prefrontal target region, located between the principal sulcus and anterior arcuate sulcus, roughly corresponding to the DLPFC or Brodmann area 46 in the human brain. Performing a decoding analysis on the recorded patterns revealed that different populations of recorded units were active in each context, and that their selectivity was highly tuned toward the task demands. More specifically, units that were active in one context adapted their firing patterns to the relevant feature dimension (for example cats vs dogs) but did not encode information about the irrelevant feature dimension. Furthermore, the selectivity to the preferred dimension was task-dependent, in the sense that the responses were attenuated in the other task where this distinction was irrelevant. These highly task-dependent signals emerged early on during the presentation period of the first stimulus

and were sustained throughout the delay. The results suggest that, as predicted by gating theories of cognitive control, DLPFC might represent the task-relevant information in a way that protects maximally against interference across tasks.

But how does information travel along the cortical hierarchy? In Siegel et al. (2015) the authors found evidence for bottom-up and top-down sweeps of information, occurring at different points of a trial. They used a similar paradigm with a bivariate stimulus space, comprised of random dots whose colour and motion direction were varied independently in four discrete steps. Monkeys were trained to categorise these stimuli either according to the colour or the motion direction, depending on the contextual cue that was shown at the beginning of each trial. In contrast to Roy et al., (2010), however, they employed a simpler paradigm in which only a single stimulus was shown on each trial, and monkeys indicated via saccadic responses whether the stimulus belonged to one or the other category. Recordings were made from a whole array of different brain regions, spanning early sensory ones such as MT, V4 and IT, which are thought to be motion, colour and object selective, and frontoparietal regions such as LIP, PFC and FEF. Quantifying the percentage of units selective for different task variables revealed that early sensory areas responded both to stimulus information (motion/colour) and cue identity, while frontoparietal regions showed stronger selectivity for task cues and choices. Interestingly, these patterns evolved dynamically over the time course of a trial. Early in the trial, the authors identified a transient burst of task information in sensory areas. This was followed by a built-up of sustained encoding of task information in frontoparietal areas, from where it propagated back to all sensorimotor areas. Crucially, they found evidence for encoding of both sensory and task-related activity across all recording sites, albeit with different temporal profiles. How were these patterns shaped by context-specific task demands? In a reanalysis of the same data, Brincat et al., (2018) compared the amount of variance within and across categories, revealing that the information represented in fronto-parietal areas was transformed into a categorical signal, while patterns in sensory areas resembled the structure present in the native stimulus space. These studies are exemplary for a large body of research spanning more than two decades, which has provided empirical

evidence for the role of PFC in task-related processing (Asaad et al., 2000; Johnston et al., 2007; Miller et al., 2002).

As alluded to earlier, advances in statistical techniques have made it now possible to investigate the geometry of these representations, i.e., how task activity is positioned in the neural state space spanned by whole populations of neurons. Intuitively, if different populations of neurons were allocated to different tasks, the resulting representation at the population level should lie on an orthogonal manifold, with separate axes of variation coding for each task. However, while such an orthogonal geometry would minimise interference in activity patterns, it would hinder generalisation between tasks. A recent study on the geometry of working memory representations for cognitive control suggests that the brain might dynamically adapt its coding schemes to prevent interference on the one hand, but also promote generalisation on the other. Panichello & Buschman (2021) trained NHPs on a retrospective and prospective cueing paradigm involving the categorisation of stimuli that varied systematically in their colour. On each trial of the retro-cue task, monkeys saw two differently coloured shapes, followed by the presentation of a task cue that indicated which of the two colours they had to recall. In the prospective cueing paradigm, the cues were presented shortly before the stimuli. This allowed the authors to distinguish between two different types of cognitive control, retroactive selection of information from working memory and prospective direction of attention to relevant features. Interestingly, they observed shared neural geometry for both paradigms in the DLPFC, providing further evidence for the domain-general involvement of PFC in cognitive control. More importantly, this geometry changed dynamically throughout a trial. First, stimulus information was encoded in two context-specific and orthogonal subspaces, which minimised potential interference between tasks. Later in the trial, however, this representation was rotated from orthogonal to parallel planes, allowing the application of the same readout mechanism to both contexts (Panichello & Buschman, 2021).

Further evidence for the existence of such parallel planes was provided by Bernardi et al. (2020), who trained monkeys to perform context-dependent lever

presses in response to various fractal images. Using cross-task decoding, in which a linear classifier was first trained on one task and then applied to the other, they found evidence for parallel representations of task variables in prefrontal areas and the hippocampus.

Task agnostic representations and nonlinear mixed selectivity

A common theme across the studies discussed in the previous section is the proposal that neural activity in PFC is highly task specific. Some evidence suggests that different populations of units are allocated to different tasks, possibly to minimise interference. Of particular appeal might be that these results could be directly explained by theories of cognitive control that proposed task-specific gating mechanisms. However, these interpretations appear to be at odds with another strand of research that reported evidence for high-dimensional and task-agnostic representations. Under this view, the coding patterns in PFC are highly idiosyncratic, mapping stimulus information into a very high-dimensional space that theoretically permits the application of various task-specific linear readouts to the same population activity. Such a coding scheme can be implemented by neurons that respond in a graded fashion to a variety of different combinations of perceptual and task variables (Fusi et al., 2016). In Rigotti et al. (2013), the authors reported evidence for such high-dimensional and task-agnostic representations in the PFC of monkeys who had been trained on a complex memory task. In this paradigm, monkeys were presented with a sequence of images that they were required to hold in working memory for subsequent recognition and recall tasks. The authors discovered that the dimensionality of neural responses was close to a mathematically estimated maximum for this task, and that this representation was implemented by neurons with non-linear mixed selectivity. While this report appears to be at odds with previous findings, it could be explained by different task demands. In contrast to the complex recognition task, cued task switching paradigms in which the same stimuli required different responses in different contexts have a larger potential for cross-task interference, which might have pressured the brain to learn a representation which minimises this cross-talk between representations (Brincat et al., 2018). Another possibility is that the dimensionality of representations, and whether they are task-agnostic

or task-specific changes across the cortical hierarchy (Ito & Murray, 2021).

1.3 Concluding Remarks

In this chapter, I explored the notion of *cognitive control* and how the prefrontal cortex might represent information and coordinate task-appropriate behaviour. The literature review highlighted that much seems to be known about the *computations* underlying task performance, and which neural substrates might be involved in task processing and error monitoring. While there appears to be consensus that PFC is involved in the representation of task-related variables, the precise geometry of these representations and how they adapt to changing task demands warrants further investigation. Furthermore, most studies to date have been carried out with NHPs. Hence, it is unclear to which extent these findings translate to human participants. Most importantly, while some of the results seem to provide evidence in favour of early models of cognitive control, these models included many hand-crafted elements. It is less clear how these gating signals might be acquired in the first place. In the next chapter, I will introduce deep artificial neural networks as toolkit to study representation learning in distributed networks of information processing units.

Chapter 2

Modelling representation learning with artificial neural networks

As mentioned in the previous section, early psychological theories relied on so-called box and arrow models which didn't specify how the proposed processes could be implemented in the neural code. In contrast, connectionist models described how networks of simple processing units, loosely inspired by biological neurons, could represent this information. However, these models usually contained many hand-crafted elements, such as the gating scheme proposed by the Miller/Cohen model of cognitive control, for which the authors manually set the context weights to specific values (Cohen et al., 1990). While research on artificial neural networks and Psychology share a common history which can be traced back to the 1950s (McCulloch & Pitts, 1943; Rosenblatt, 1958), fruitful interactions were severely hindered by the lack of sufficient compute and growing disinterest in artificial neural networks among AI researchers in the second half of the last century (Minsky & Papert, 1969). However, over the past two decades, changes in the affordability and overall availability of massive compute have led to a resurgence of interest in artificial neural networks for engineering applications, which is often referred to as the Deep Learning revolution (LeCun et al., 2015). As the name suggests, deep learning involves very deep artificial neural networks, which are stacked layers of non-linear function approximators, trained end-to-end via sophisticated optimisation procedures (LeCun et al., 2015). It may not be immediately apparent why these advances could be relevant to neuroscientists.

However, neuroscience itself underwent a transformation in the last two decades, as the focus shifted slowly from single-neuron studies to analyses of the geometry and dimensionality of activity embedded in populations of neurons (Saxena & Cunningham, 2019). Artificial neural networks trained end-to-end on cognitive tasks provide a testbed for theories of representation learning, as they too consist of numerous simple processing units which are interconnected to form a very powerful universal function approximator (Yuste, 2015). Indeed, when trained to categorise a sufficiently large and heterogenous dataset of images depicting naturalistic scenes and objects, individual units (more precisely filters) in early layers in a particular class of feed-forward neural networks, so-called convolutional neural networks, evolve shape and colour sensitive tuning profiles not unlike those observed in the early visual cortex of the mammalian brain (Lindsay, 2021). These early observations inspired a fundamental change in the way cognitive/computational neuroscience is carried out, with many researchers now adapting a deep learning framework to study how task representations might emerge in various areas of the human brain (Lindsay, 2021; Richards et al., 2019; Saxe et al., 2021; Storrs & Kriegeskorte, 2019), In this section, I will argue that the success of deep neural networks as models of ventral stream processing suggest promising avenues for research on representation learning in cognitive neuroscience, but that this exercise should be carried out with caution. I will begin with a quick primer on artificial neural networks and their training methods, followed by a short review of studies that have revealed parallels between visual information processing in biological and artificial neural networks. I will then discuss some key differences between brains and machines and highlight the utility of a much simpler class of models. In doing so, I hope to convey the idea that neural networks are promising models of how representations are sculpted by experience, but further work is required to understand how this may or may not relate to human learning.

2.1 A primer on artificial neural networks and deep learning

Artificial neural networks are highly non-linear function approximators (Goodfellow et al., 2016). In this thesis, we will focus on supervised learning, which involves learning generalisable mappings from inputs to outputs, as for example from images to labels that denote the class membership of each image.

A typical neural network consists of an input layer, one or more hidden layers and an output layer (Goodfellow et al., 2016). Each hidden layer consists of several units, each of which represents a linear transformation of the information encoded in the previous layer, multiplied with a weight vector and added to a bias term (also known as parameters). Commonly, the outputs of these units are passed through a non-linear activation function. Depending on the choice of these units, the deep neural network is theoretically able to approximate any arbitrary function (Cybenko, 1989). To approximate a function with a neural network, its parameters are adjusted with a training procedure that minimises a loss, or objective function. We will begin with a brief overview of the history of neural networks, to highlight early interactions between AI research and Psychology. I will then explain how neural networks are trained. Lastly, I will introduce a particular class of neural networks that has been successfully applied to computer vision problems, called convolutional neural networks.

2.1.1 A brief history of artificial neural networks

The foundations for modern neural networks were already laid in 1943, when McCulloch and Pitts introduced elementary units with binary response profiles as abstract models of biological neurons (McCulloch & Pitts, 1943). The authors didn't address the question of how connections between these units could be learned. A solution was proposed less than two decades later, when Frank Rosenblatt introduced the Perceptron, a function whose binary response depended on whether its weighted input exceeded a certain threshold or not. He proposed a learning algorithm that would increment the current task weight based on the mismatch between the expected and produced output, multiplied by the corresponding input (Rosenblatt, 1958). While the

perceptron received considerable attention initially, as it could be used to learn solutions to binary classification problems and was, even if only very remotely, inspired by biological neurons, much of the enthusiasm surrounding artificial neural networks was dampened when Minsky and Papert published their famous critique of the perceptron algorithm in 1969 (Minsky & Papert, 1969). They argued that a single perceptron unit with two input nodes was unable to return one if both inputs were unequal and zero if they were equal, which is known as the logical XOR-problem. While they acknowledged that this could be solved by introducing multiple layers of perceptron units, methods to efficiently train such a network didn't exist until the 80s, when the backpropagation algorithm, initially introduced by Werbos (1974), slowly gained mainstream popularity (Rumelhart et al., 1986). In the next section, I will discuss procedures to train deep neural networks and introduce more contemporary methods that have been successfully applied to problems in neuroscience.

2.1.2 Training procedure

The optimisation procedure that allows a neural network to express the functional relationship between inputs and labels of a dataset is called *training*. In this section, we review the different components of this training procedure.

Loss functions

Loss functions quantify the extent to which the neural network's outputs match the ground-truth and are essential for supervised learning. The choice of loss function is dependent on the type of data and machine learning problem. For supervised learning, one can broadly distinguish between regression and classification problems, with continuous and categorical outputs respectively. For regression problems, the loss is defined as mean squared error between the true label and the network's prediction:

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}(x_i;\theta))^2 \tag{2.1}$$

where θ represents the network parameters, y the true label and $\hat{y}$ the network's prediction (Goodfellow et al., 2016).

In contrast, for classification problems, the output from the last linear unit is

passed through a nonlinearity, such as a sigmoid for binary classification or SoftMax for multi-class problems. To train such a network, the cross-entropy loss function is used, which captures the inefficiency with which the network's output describe the ground-truth labels (Goodfellow et al., 2016):

$$L(\theta) = -\sum_{i=1}^{m} y_i log(\hat{y}(x_i;\theta)) \tag{2.2}$$

Stochastic Gradient Descent

At the beginning of the training procedure, neural networks are initialised with random parameters. To find the values that minimise the loss function, iterative updates are applied to all network parameters in the direction of the gradient of the loss with respect to the network's parameters. This optimisation procedure is called *Gradient Descent* and works as follows. Given a loss function $L(\theta) = \frac{1}{m}\sum_{i=1}^{m} l_i(\theta)$ where $l_i(\theta)$ corresponds to the loss of a single datum in m, the goal is to change θ so that $L(\theta)$ is minimised.

This requires the first derivative of the loss function with respect to the network parameters θ. For a vector-valued function, this derivative is the gradient $\nabla^{(\theta)}L(\theta)$ over the input weight vector (Goodfellow et al., 2016).
To reduce the loss, the parameters are updated by subtracting this gradient, scaled by the learning rate η, from their current values:

$$\theta_{t+1} \leftarrow \theta_t - \eta\nabla^{(\theta)}L(\theta) \tag{2.3}$$

This process is repeated until convergence. In practice however, the gradient is usually computed for a subset of the available training data, which is called *stochastic gradient descent*: In the most extreme case, the gradient is evaluated *online* on a single sample (Goodfellow et al., 2016):

$$\theta_{t+1} \leftarrow \theta_t - \eta\nabla^{(\theta)}l_i(\theta) \tag{2.4}$$

As individual samples can be quite noisy, so-called *minibatches* are usually used

which represent random subsets S_j that are sampled from the training set and used to compute the gradient.

$$\theta_{t+1} \leftarrow \theta_t - \eta \sum_{i \in S_j} \nabla^{(\theta)} j_i(\theta) \tag{2.5}$$

The Backpropagation algorithm

A neural network consists of several layers of units, each of which has their own weights and biases. How can all these parameters be updated efficiently? To apply SGD, one needs to calculate the partial derivative of the loss with respect to each parameter. A naïve approach would be to begin at the input and progress to the output, computing the partial derivate of each unit with respect to its input. Then, by applying the (multivariate) chain rule of calculus, one could obtain the partial derivative of the loss w.r.t. each parameter by multiplying all partial derivatives along each path from the parameter of interest to the output, and then summing up these paths. The so-called backpropagation algorithm, developed in the 1970s, renders this problem tractable thanks to a more efficient application of the multivariate chain rule of calculus (Rumelhart et al., 1986; Werbos, 1974). The network is traversed starting from the output. For each node, partial derivatives of all paths originating from that node are summed up, yielding partial derivatives of the output with respect to that node. As partial derivatives of the output with respect to early layers include partial derivatives of later layers, this prevents redundant computations and reduces the complexity of the differentiation procedure substantially.

2.1.3 Efficient deep learning with convolutional neural networks

Standard feed-forward neural networks are *fully connected*. Each unit in a layer receives inputs from all units in the preceding layer and propagates information forward to all units of the following layer. In particular for computer vision applications, this introduces some serious problems with scalability. Take for example an RGB image of 128x128x3 pixels. Each unit in the first hidden layer would then receive 128x128x3 = 49152 inputs, each multiplied with a separate network weight. For a network with 100 hidden units, the first hidden layer alone would then already have 4.9 million parameters. Leaving scalability aside, another problem for computer vision applications

is that objects of interest usually only occlude a small part of the input image (locality) and, for natural scenes, can occur in different locations of the image (translation invariance). A fully connected network does not exploit these image properties.

Convolutional Layers

Convolutional Neural Networks (CNNs or ConvNets) have been developed to capitalise on the locality and translation invariance properties of naturalistic images (Fukushima, 1980; Lecun et al., 1998). A *convolutional layer* consists of several *convolutional kernels*, or *filters* which are convolved with the input image. Convolutional layers are often illustrated as cubes, with the first two dimensions corresponding to the kernel size and the third dimension corresponding to the number of kernels. In contrast to the flattened inputs passed into a standard feed-forward network, ConvNets are able to operate on the original two-dimensional input image. Each kernel is moved progressively across the input image, and the output at each location is given by the dot product between the filter and the receptive field in the input image, creating a so-called *feature map*. The size of this feature map is given by $M = N - k + 1$, where M is the size of the output, N input size and k the kernel size. The feature maps are usually passed through a ReLU non-linearity. Stacks of several convolutional layers form a convolutional neural network (LeCun et al., 1998).

Pooling Layers

Depending on the size of the inputs and number of kernels, the output of a convolutional layer might still be quite high-dimensional. So-called *pooling layers* can be used to down sample the layer output. The most common technique is max-pooling, which returns the unit with highest activity within a receptive field (Goodfellow et al., 2016). In a CNN convolution and max-pooling layers are usually alternated several times, before the output is flattened and passed through one or two fully connected layers and a final output layer (LeCun et al., 1998).

2.2 Deep neural networks for neuroscience

The past decade has been marked by unprecedented advances in deep learning research, in particular in the fields of computer vision and natural language processing (Voulodimos et al., 2018; Young et al., 2018). Over the timespan of only a couple of years, neural network models have been refined to a state where they routinely outperform even human experts on computer vision tasks such as object recognition (Voulodimos et al., 2018). These advances in engineering have caught the attention of neuroscientists, who have begun to investigate parallels between representations learned by these models, predominantly convolutional neural networks trained with a supervised objective, and those observed in the mammalian brain (Lindsay, 2021; Richards et al., 2019). In this section, I will review research on commonalities between deep neural networks and the mammalian brain, both in terms of the learned representations and the dynamics of the learning process itself. I will then discuss differences between minds and machines and argue that much simpler classes of models are needed, to prevent replacing one black box by another.

2.2.1 Representational equivalence between brains and machines

Most of the early work on commonalities between brains and neural networks has focussed on vision (Saxe et al., 2021; Lindsay, 2021). In their seminal paper, Khaligh-Razavi & Kriegeskorte (2014) investigated whether supervised or unsupervised computer vision models could explain activity patterns recorded from human (fMRI) and macaque (single-cell recordings) brains. Using Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), they compared activity patterns in early visual areas and IT to different layers of a convolutional neural network and various other computer vision models trained on image classification tasks. Crucially, among these models the CNN performed best in explaining neural patterns. Within the CNN, recordings from IT correlated strongest with deep layers, while representations in early layers of the CNN were highly similar to those observed in early visual areas of the human and NHP brain, which was interpreted by the authors as evidence that ventral stream visual processing has a similar hierarchical organisation as deep neural networks trained on classification tasks (Khaligh-Razavi & Kriegeskorte, 2014). Similar results were re-

ported by Yamins et al. (2014) who discovered that the final categorisation accuracy of a model trained on object classification determined how well its representations matched those observed in area V4 and IT of the macaque brain. These findings are particularly striking as in both cases, models were optimised for task performance, rather than to predict neural activity patterns.

Crucially, these striking similarities between minds and machines are not limited to vision. Chaisangmongkon et al. (2017) trained recurrent neural networks (RNN) and NHPs on a delayed match to category task to investigate how representations of task variables evolve over the time course of a trial. Interestingly, both in the RNN and in area LIP and PFC of the monkey brain, they observed transformations from representations that distinguished between different stimulus types early on to a binary variable that encoded the response in later time points of the trial, suggesting that neural networks which are trained end to end on cognitive tasks might serve as computational models for higher order information processing (Chaisangmongkon et al., 2017; Engel et al., 2015).

2.2.2 Differences between computations in brains and machines

It may appear obvious that deep neural networks trained with back-propagation can only serve as very crude approximation of biological information processing (Lillicrap et al., 2016), but even if one is content with the view that learning can be characterised as loss function minimisation via weight changes in distributed networks of non-linear functions (Richards et al., 2019), limitations of this approach must be acknowledged. How exactly a neural network arrives at a particular solution is only poorly understood, which implies that using neural networks as models of the brain could potentially lead to replacing one black box by another (Saxe et al., 2021). Indeed, some evidence suggests that state-of-the-art deep neural networks and humans might process information in fundamentally different ways (Bowers et al., 2022). In this section, I will briefly review some findings in the domain of computer vision to motivate the use of simple, hence tractable models, and discuss the fundamental issue of catastrophic forgetting, which is the inability of current neural networks to learn multiple tasks in succession.

Visual information processing

The benefit of RSA is that it is agnostic to the process that gave rise to the observed representation, which means that the technique can be used to compare different recording modalities, patterns in artificial and biological networks and neural activity to behaviour (Kriegeskorte et al., 2008). However, this approach provides only a limited insight into the computations that underlie the formation of the observed representations. That this methodological limitation might pose the danger of overstating equivalences between brains and machines has recently been exemplified by research on the inner workings of trained CNNs. The most prominent shortcoming of CNNs is their vulnerability to adversarial attacks. That is, adding certain types of noise to input images is sufficient to change the output of the network, so that it may for instance classify a noisy image of a cat as a caterpillar, while the performance of human participants on the same task remains the same (Akhtar & Mian, 2018). Follow-up work suggests that humans and deep neural networks indeed appear to process visual information in fundamentally different ways (Dujmović et al., 2020; Geirhos et al., 2020). First, CNNs seem to have a texture bias, while humans rely on the overall shape of objects for classification judgements. Geirhos et al. (2019) demonstrated this with a dataset in which the textures of objects were systematically altered, so that for instance a dog would be depicted with elephant skin, While humans would still be able to classify the wrinkly dog as dog, CNNs erroneously classified it as elephant. Another issue is that CNNs classify based on local, instead of global structure. Baker et al. (2018) trained CNNs on contour images and evaluated it on a dataset in which some images had fuzzy instead of smooth edges (local perturbation) and others where large image patches were randomly permuted (global perturbation). Interestingly, the performance of CNNs was strongly impaired by local perturbations, which, again had only a marginal impact on human categorisation performance.

While the extent to which patterns in different layers of a CNN can predict those recorded from brains seems striking, and previous studies have shown that this scales with the performance a network achieves on a classification task, a surprisingly small fraction of explained variance can be attributed to the training procedure. For instance,

Güçlü et al. compared correlations between neural patterns and CNNs that were either randomly initialised or trained to convergence on a classification task. Surprisingly, for most of visual cortex, the accuracy with which the randomly initialised network could predict BOLD patterns was just 5-20% below the one of the fully trained network (Güçlü & Gerven, 2015). It is less clear how this might differ for higher cortical areas that are engaged in more task-specific processes.

Taken together, these examples suggest that merely correlating the outputs of individual layers of a network and recordings from different brain areas might be insufficient to understand biological information processing, as the same type of output can potentially be produced by vastly different underlying computations.

Catastrophic forgetting

Another fundamental issue is continual learning. In contrast to humans, who can learn multiple tasks over their lifetime, training a neural network on a new task usually leads to so-called *catastrophic forgetting* or *catastrophic interference* (French, 1999). To model the dynamics of multi-task learning in humans with neural networks, this limitation must be overcome. Finding biologically inspired solutions to continual learning is an unsolved research problem, predominantly as the mechanistic underpinnings of continual learning in humans are also only poorly understood (Parisi et al., 2019; Hadsell et al., 2020). A very simple solution would be to randomly interleave training data from multiple tasks, but humans seem to perform worse under such a curriculum (Flesch et al., 2018).

2.2.3 Understanding and overcoming limitations of artificial neural networks

If state-of-the-art neural networks appear to learn in ways that are potentially quite different to biological brains, how could they still be used as models of human information processing? One possible soltution would be to turn to highly simplified model architectures first, and study their training dynamics and how they arrive at their solutions, to derive principled prediction of neural information processing which could then be

tested in biological circuits, rather than correlating the activity of large off-the-shelf models and brain areas directly (Saxe et al., 2021). In this section, I will introduce two methods from deep learning theory which have provided interesting insights into the way in which neural networks learn, and argue that simplified models could be used as mathematical toolkit to form general predictions for questions in neuroscience. I will then turn to the problem of catastrophic forgetting. I will argue that the reasons of catastrophic forgetting are very well understood, thanks to the use of very simple models, and that this has inspired a variety of engineering solutions. A major limitation, however, is that current solutions are only loosely inspired by the brain, which suggests avenues for further research.

Understanding learning dynamics

Precisely characterising the learning dynamics of a model with millions of parameters to understand how and why certain computational motifs arise seems like an impossible feat. For this reason, deep learning theoreticians have turned to much simpler classes of feed-forward models, sometimes even without non-linearities, as their dynamics are much more tractable and can be scrutinised with a whole array of powerful mathematical techniques.

Deep linear networks . For deep linear models, it is possible to derive analytic solutions to the training dynamics (Saxe, 2015). This has led to interesting insights into the way in which neural networks acquire information. For example, psychological studies suggest that humans first acquire broad distinctions between object classes, which are then progressively refined, such as when first learning about cats and dogs, and later about Beagles and Golden Retrievers (Carey, 2011). While deep neural networks are known to exhibit similar learning dynamics, only by analysing deep linear networks, it has become possible to understand why they behave in that way. Saxe et al. (2015) reported that during training, the network arrives at saddle points in the loss surface, leading to stage-like transitions. Applying Singular-Value-Decomposition to the input-output correlation matrices of the network revealed that this progressive differentiation is driven by the way in which these correlations impact learning over time, so that larger components are learned prior to smaller ones. Future work could in-

vestigate how these findings generalise to more complex models, and which specific computational motifs might lead to divergent findings in brains and machines.

The Neural Tangent Kernel. Another line of research in deep learning theory investigates the behaviour of networks in the infinite width limit. Interestingly, in this limit, learning dynamics can drastically simplify, which makes it possible to derive analytical solutions even for neural networks with non-linear activation functions. The Neural-Tangent-Kernel (NTK) theory provides such an analytic solution to the learning dynamics in this limit (Jacot et al., 2020; Lee et al., 2019; Sohl-Dickstein et al., 2020). It is based on the observation that functions expressed by a randomly initialised network can be described as draws from a Gaussian Process (Lee et al., 2018). Intuitively, this makes sense as neural networks are initialised with draws from Gaussian distributions. The sum of all inputs to a unit is a sum of Gaussian random variables, which itself follows a normal distribution (Central Limit Theorem). While the equivalence between neural networks and Gaussian Processes only holds at initialisation, authors of the NTK theory observed that in the infinite width limit, this also holds throughout the training process (Jacot et al., 2020). The reader might wonder how this applies to practical applications of neural networks, which do not operate in an infinte-width regime. Also in 2020, Chizat and colleagues discovered that any neural network can be pushed in this regime by initialising its parameters with sufficiently large values (Chizat et al., 2020). They report that the scale of the variance of weights at initialisation determines whether the network operates in a "rich" regime, where its parameters change substantially over time, or in a "lazy" regime, where its dynamics can be described by the NTK (Chizat et al., 2020; Woodworth et al., 2020). From a neuroscience perspective, this is interesting as the two regimes appear to correspond roughly to task-specific versus task-agnostic representation learning and hence might provide a toolkit to explore how different representational geometries emerge and affect task performance, a research avenue I will explore in detail in **Chapter 4**.

Together, these two examples show how deep learning theory might be able to shed light on the inner workings of neural networks, and how turning away from large off-the-shelf architectures could help neuroscientists to derive principled prediction of

how representations could evolve over time in biological circuits.

Understanding and overcoming catastrophic forgetting

Another obstacle for neuroscientists is the inability of artificial neural networks to learn continually without forgetting. One of the first investigations of this phenomenon of catastrophic forgetting was carried out by McCloskey & Cohen (1989), who contrasted the effect of continual ("blocked") and interleaved training curricula on the final performance and error patterns of simple feed-forward neural networks with three layers and sigmoidal non-linearities.While the network was able to learn two tasks under interleaved training, it forgot how to perform the first task after training on the second under blocked training. Disruption of previous knowledge has also been observed in human participants, as for instance in (Barnes & Underwood, 1959), who reported so-called *retrograde interference* in a task that involved learning pairwise stimulus-response associations. McCloskey and Cohen trained neural networks on this Barnes & Underwood paradigm. In contrast to human participants, the network completely forgot the associations learned first, afterit was trained on associations with novel stimulus pairs. The authors speculated that catastrophic interference happened as representations in the neural network were highly distributed. To illustrate this, they introduced the concept of a high-dimensional "weight space", spanned by the network parameters. They referred to regions in this space that minimised the loss for a given task as "solution space". The amount of forgetting should then depend on the extent to which solution spaces between different tasks are shared, and whether the network is steered to a position in this space that minimises the objective for both tasks. The authors attempted to mitigate forgetting by changing the learning rate, training duration and even overall number of network parameters, but none of these measures was sufficient to prevent catastrophic interference (McCloskey & Cohen, 1989).

To summarise, catastrophic forgetting appears to be an inherent property of neural networks that have been trained with gradient descent to learn distributed task representations. It is thought to occur whenever network parameters change to adapt to novel task demands (French, 1999). As the problem is very well characterised, many attempts have been made to change the way in which neural networks learn so that their

weight configurations remain in the solution space shared by multiple, successively learned tasks. These solutions can be broadly divided into four different groups (see also Parisi et al., 2019 and Hadsell et al., 2020). (1) Regularisation approaches augment the loss function so that the network is encouraged to retain information about previously learned tasks. (2) Dynamic architecture growth approaches increase the capacity of the neural network progressively, either by adding columns or layers with each new task that is encountered. (3) replay approaches continually sample previously experienced trials, either from a storage buffer or a generative model, to artificially interleave the training procedure. (4) Lastly, orthogonalization approaches push task updates into directions that are orthogonal to representations of previously acquired tasks, either by explicitly gating hidden units or by changing the direction of the gradient updates. Before I continue with a review of these approaches, it should be noted that the continual learning problem is multi-faceted and no universally applicable solution might exist (Parisi et al., 2019; van de Ven & Tolias, 2019). In this thesis, I focus on a specific type of continual learning, which is commonly referred to as task incremental learning (van de Ven & Tolias, 2019). In this scenario, the network is required to continually learn novel tasks for the same stimulus set and is provided with an explicit task cue. This can be contrasted with other scenarios such as class incremental learning, where additional categories are introduced, such as when a network first learns to categorise ones and twos, and later how to classify fours and fives (van de Ven & Tolias, 2019).

Regularisation approaches: Catastrophic Interference seems to occur when different tasks require different configurations of task weights (French, 1999). However, modern deep neural networks are usually significantly overparameterized for the task at hand (Sejnowski, 2020) which could suggest that they might have substantial free capacity to learn additional tasks (Gou et al., 2021). Regularisation approaches such as Elastic Weight Consolidation (EWC) and Synaptic Intelligence (SI) rely on an augmented loss function that penalises the network for changing weights that are deemed important for previously learned tasks, hence essentially partitioning the network into subnetworks with task-unique weight configurations. For EWC, the supervised loss is augmented with a regularisation term that penalises squared deviations between the

weights learned for a previous task and the current task weights, where each weight's contribution to the loss is scaled by its Fisher information, which quantifies the extent to which the solution learned for the previous task relied on a particular weight (Kirkpatrick et al., 2017). Mathematically, the Fisher information measures the "peakiness" of the log-likelihood of the task data conditioned on the network weights. If changes in the weight are accompanied by large changes in the log-likelihood, the specific value of that weight is crucial for that task. For two tasks A and B, the regulariser has the form

$$L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \tag{2.6}$$

where θ^*_A are the optimal parameters for task A, λ is a scaling hyperparameter and F denotes the Fisher information matrix (Kirkpatrick et al., 2017).

A very similar approach was proposed by Zenke et al, called Synaptic Intelligence (SI, (Zenke et al., 2017)), who compute the importance of a weight based on how it has changed over the time course of the training phase. Like EWC, they penalise squared deviations of the current task weights from those learned for the previous task:

$$L(\theta) = L_B(\theta) + c \sum_i \Omega^n_i (\theta^*_i - \theta)^2 \tag{2.7}$$

where Ω denotes the parameter importance.

From a neuroscience perspective, an appealing property of these regularisation approaches is that they dynamically change the plasticity of neural changes depending on their importance for continual task performance. Similar behaviour has been observed in neurons of biological brains which means that these approaches are at least partially biologically plausible (Benna & Fusi, 2016). A major drawback, however, is that to compute the regularisation term, a copy of the weights that had been learned for the previous task must be kept in memory. If it's possible to maintain a copy of these weights, one might wonder why one couldn't simply store a copy of the old network, instead of performing these regularised weight updates.

Replay approaches: Another type of methods which builds on findings from neuroscience relies on replay techniques. The idea that previous experiences are replayed

from memory during learning is central to the so-called complementary learning systems (CLS) theory. Under this theory, the Hippocampus stores snapshots of experiences, while more abstract, semantic memory representations are encoded in the neocortex (McClelland et al., 1995). Learning involves a gradual transformation from concrete – or *episodic* - to abstract semantic representations, which is aided by continuous replay of previous experiences from the hippocampal memory. Numerous variants of these replay mechanisms have been proposed as solutions to continual learning. The simplest implementation relies on a separate memory buffer. Throughout training, pairs of inputs and labels are occasionally transferred to and randomly sampled from this buffer to artificially interleave the training procedure (Rolnick et al., 2019). Storing large quantities of training examples might not be practical for contemporary deep learning approaches. Alternative methods rely on so-called *pseudo-rehearsal*, where a generative model of the data is learned from which training data can be sampled ad-hoc (Shin et al., 2017; van de Ven et al., 2020). While replay methods are motivated by a neuroscientific theory, it's unclear which training samples should be stored and how this storage mechanism could be implemented in a biologically plausible manner.

Dynamic architecture growth: Early research on the underlying causes of catastrophic forgetting suggested that it is an inherent property of networks that acquire highly distributed representations, as each weight is involved in the representation of a task to some extent, and changes to the network's overall weight configuration hence are likely to disrupt previously encoded memory. This suggests that a simple solution would be to encourage sparseness in the neural representations, whereby different units are allocated to different tasks. Progressive neural networks make this explicit by progressively adding new columns to the network with each novel task it encounters (Rusu et al., 2016). First, a standard feed-forward network is trained on one task. The authors refer to this initial network as "column". Then, as a new task is encountered, a new column is added to the network, weights of the first column are frozen to prevent them from changing and lateral connections between the first and second column are introduced to promote reuse of existing feature representations. To make this scalable, the authors perform dimensionality reduction on the lateral connections, so that

only n inputs are provided to each column, irrespective of the number of learned tasks. Practically, this is implemented with an additional small-scale neural network of fixed hidden layer size n, which the authors refer to as *adapter*. While the approach seems appealing due to its architectural simplicity, it has several major limitations. Most importantly, progressive neural networks don't scale well, as one column is added per new task. Secondly, as past task columns are exempted from training, the network might have a lot of redundant capacity, in particular if tasks are similar to each other.

Orthogonalisation and gating approaches: As discussed in the previous section, Catastrophic Forgetting could potentially be prevented with sparse representations, where different tasks lie in different, ideally orthogonal locations in the neural state space. Orthogonalisation approaches attempt to bias the network towards learning such orthogonal task representations. Broadly speaking, these can be divided into approaches that act on the gradient updates and those that implement a form of gating on the individual units of a hidden layer. An example of the former has been proposed by (Zeng et al., 2019), who introduced the Orthogonal Weight Modification (OWM) algorithm. Once the network has been trained on the first task, OWM implements a projection matrix that moves gradient updates for the next task into directions that are orthogonal to the space spanned by the inputs for the previous task:

$$P_l^j = I - A_l(A_l^T A_l + \alpha I)^{-1} A_l^I \tag{2.8}$$

for task j and layer l, where A contains all previous inputs to layer l. The gradient in layer l is multiplied with this projection matrix before the standard SGD update is performed:

$$W^l \leftarrow W^l - \varepsilon P_l^B \nabla W^l \tag{2.9}$$

To compute this projection matrix, the algorithm requires access to the entire training data from the previous task. An alternative formulation of the OWM algorithm performs online updates using the Recursive Least Squares (RLS) algorithm, which

updates the projection matrix after each training sample/minibatch n:

$$P^{RLS}(n) = \alpha^{-1}[I - A(\alpha I + A^T A)^{-1} A^T] \tag{2.10}$$

A similar technique was proposed by (Farajtabar et al., 2019) who suggested to project gradient updates into an orthogonal subspace where the outputs for the previous task don't change. From a neuroscience perspective, it might be unclear how these orthogonalization approaches could be implemented in biological circuits. An alternative method that orthogonalizes hidden layer representations relies on a gating strategy, which can be related to early theories on prefrontal cortex function discussed earlier. For example, (Masse et al., 2018) introduced a mask that selectively activates non-overlapping subsets of units for each learned task. A similar approach was recently suggested by (Russin et al., 2022) wo apply a multiplicative mask to hidden neurons to partition the layer into task-specific subsets. These gating approaches achieve orthogonalization via *axis alignment* of hidden layer representations: As different banks of units are allocated to different tasks, any crosstalk between these tasks can be fully prevented. A major drawback of the work to date is that these gating schemes are usually hand-crafted by the experimenter. While they are directly motivated by theories on Cognitive Control and prefrontal cortex function, follow-up work would need to develop a biologically plausible mechanism that can learn these gating strategies from scratch.

2.3 Concluding remarks and aims of thesis

The past decade has been marked by a significant change in the way cognitive neuroscience research is carried out. Predominantly in vision, researchers have embraced a deep-learning approach, where large-scale neural networks are trained on similar stimuli and/or tasks as human participants and serve as models for the types of representations observed in different regions along the ventral stream of the mammalian brain. Similar correspondences between artificial and biological neural networks have been observed in the domain of task learning and decision-making, suggesting that

these architectures might serve as testbed for theories about neural computation. This enthusiasm, however, has been hampered by more recent work which revealed that CNNs and brains process information in fundamentally different ways, while the precise reasons for these differences remain to be understood. A potential solution to this problem is to start with much smaller, hence tractable neural network models which are similar to those used by connectionists in the 1980s, but don't include any hand-crafted elements. Insights from deep learning theory might help to understand how representations are learned in these "toy models", and should precede experiments with more complex architectures to study the observed effects at scale (Saxe et al., 2021). Another issue is that neural networks fail to learn continually without catastrophic forgetting. Interestingly, the reason for this phenomenon seems very well understood, which has led to various proposals on how to prevent these catastrophic forms of forgetting. From a neuroscience perspective, however, many of these solutions lack biological plausibility. More importantly, they don't seem to capture the apparent benefit of blocked (or continuous) over interleaved training regimes, which have been reported in certain variants of category learning problems in humans (Carvalho & Goldstone, 2014; Flesch et al., 2018).

Having discussed the literature on cognitive control, task representations and deep learning approaches to cognitive neuroscience, we can now revisit the main question of this thesis. What are the computations and representations that permit continual task performance? Previous research has predominantly focussed on the control processes that allow the brain to change swiftly between different context-appropriate processing patterns. Of particular appeal are theories that prescribe some form of gating to prefrontal cortex, and some evidence from neuroscience seems to support this view. However, the research on the neural representations of tasks seems relatively inconclusive. First, evidence points either towards task-specific representations, as predicted by the gating theory, or to much more general task agnostic representations and non-linear mixed selectivity. Secondly, there seems to be a general lack of neuroscientific studies with human participants. Thirdly, a theoretical investigation of the conditions under which these processing schemes might emerge appears to be lacking in the lit-

erature. While deep learning models have been used extensively as models of visual processing, their potential to study multi-task learning remains largely unexplored and is hindered by their inability to learn continually without forgetting. Moreover, I have reasoned that applying off-the-shelf models poses the risk of replacing one black-box by another, and argued in favour of using small, tractable models to gain insights into the computational building blocks. Lastly, from a cognitive perspective, most studies on the neural geometry of tasks seem to have focussed on simple stimuli with perceptually-varying feature dimensions. A hallmark of human cognition, however, is the ability to learn abstract representations, which can be applied across different domains – such as for instance the general concept of "dangerousness" or "size". In this thesis, I attempt to tackle these gaps in the literature with a series of computational, neuroimaging and behavioural investigations of the key properties of continual representation learning.

More specifically, I seek to (a) provide a computational theory of how information for context-dependent decision making could be represented, (b) test predictions from this theory in recordings from the human brain, (c) revise and extend my computational model to overcome the problem of catastrophic forgetting, and (d), test whether humans similarly benefit from continual, in contrast to interleaved training, to learn abstract, generalisable concepts. In the following, I describe how these topics will be addressed in the remaining chapters of this thesis.

In **Chapter 3** of this thesis, I will present a computational theory of how representations are sculpted for context-dependent task performance. Using small feed-forward neural networks, I will explore how different training modes, dubbed "rich" and "lazy" learning, affect what and how a neural network learns to perform context-dependent decision making tasks. I will investigate how the representations are formed from a computational perspective, and how these findings relate to earlier work on cognitive control, in particular theories on gating in PFC. The chapter shall end with a simulation involving state of the art Convolutional Neural Networks, to explore how these findings generalise to more complex architectures.

Having explored the space of different neural geometries, will I present an empir-

ical study in **Chapter 4**, where we trained human participants on a context-dependent decision making task and recorded fMRI data during a subsequent test phase. This is sought to fill a gap in the literature on task representations, which has predominantly focussed on NHPs and much simpler tasks. I will also present results from a supplementary supplementary re-analysis of single-cell recordings from the macaque brain that allowed me to test more fine-grained predictions from the previous chapter.

In **Chapter 5**, I will present a novel training method for feed-forward neural networks which captures key properties of human continual learning. More precisely, the model will capture the benefit of blocked over interleaved training which has been previously observed (Flesch et al., 2018).

Much of the previous work has focussed on simple cognitive tasks with perceptually varying stimuli. In **Chapter 6**, I will explore the possibility that blocked, in contrast to interleaved learning, allows humans to learn abstract mappings from stimuli to responses, such as whether "dangerousness" or "speed" determines whether an object should be chosen, and that these can be generalised to stimuli from other domains that share the same characteristics.

This thesis ends with a general conclusion, which revisits the key findings and discusses key limitations as well as suggestions for future work.

Chapter 3

Rich and lazy learning of task representations in neural networks

Abstract

The neural coding scheme that permits continual task performance in humans is only poorly understood. Recent advances in deep learning have led to a surge in interest in the application of deep artificial neural networks as models for neural information processing. Here, we demonstrate that this exercise should be carried out with caution, as the same architecture can learn strikingly different representations, depending on the way in which it is initialised. Using insights from deep learning theory, we demonstrate how a context-dependent decision making problem can be solved with two different regimes, called "rich" and "lazy" learning, which trade-off learning speed for robustness, and lead to the formation of either highly task-specific, low-dimensional and orthogonal, or task-agnostic and high-dimensional representations. In subsequent simulations with deeper neural networks, we demonstrate that under rich learning, information is transformed in several processing stages from a format that resembles the structure of the input space, to highly task-specific representations in the reference frame of the responses. Together, our simulations provide a computational framework to characterise the range of coding schemes for context-dependent processing, and lay the foundation for computationally-informed hypotheses about the format in which task information is encoded in the human brain.

3.1 Introduction

How can the brain code for multiple, potentially conflicting tasks? While previous work has extensively studied the control processes underlying successful task performance, much less is known about the format in which task information is represented in the brain, and how these representations are sculpted when novel tasks are learned (Badre et al., 2021; Miller & Cohen, 2001). Here we introduce a computational framework to study how neural networks solve context-dependent decision-making tasks and derive testable predictions for the error patterns and representational geometries that might emerge in the brains of human participants trained on comparable tasks.

An emergent theme in machine learning research is that neural networks can solve non-linear problems in two distinct ways, dubbed the lazy and rich regimes, which respectively give rise to high- and low-dimensional representational patterns in the network hidden units (Arora et al., 2019; Chizat et al., 2020; Jacot et al., 2020; Lee et al., 2019; Woodworth et al., 2020). In the lazy regime, which occurs when weights in the hidden layers are initialised with draws from a distribution with high variance, the dimensionality of the input signals is expanded via random projections to the hidden layer, such that learning is mostly confined to the readout weights. In the rich regime, which occurs under low initial variance, the hidden units instead learn highly structured representations that are tailored to the task demands (Geiger et al., 2020; Paccolat et al., 2021; Saxe et al., 2019; Woodworth et al., 2020).

We used neural network simulations to characterise the nature of these solutions for a canonical context-dependent decision-making setting and employed representational similarity analysis to explore their neural geometry. Rich and lazy learning induces different representational geometries, while trading off learning-speed for robustness. While lazy networks were faster to converge, their outputs were stronger affected by the addition of random noise to the inputs. The solution learned under rich learning involved representing distinct tasks as low-dimensional and task-specific neural manifolds, in a way that minimises interference and maximises robustness among potentially competing responses (Koay et al., 2019). Task-relevant features were mapped onto orthogonal dimensions in neural state space. Relative to these,

task-irrelevant features were strongly attenuated. In contrast, under lazy learning, the network recapitulated the structure of the stimulus space in a task-agnostic manner. To bridge the gap between these toy models and modern architectures, we confirmed that very similar patterns emerge in a convolutional neural network trained directly on high-dimensional visual stimuli from a previous experiment with human participants. Furthermore, under rich learning in this deep network, we found evidence for a successive transformation of hidden layer representations from the input space to a generalisable encoding of the responses.

3.2 Results

3.2.1 The initial weight scale of a neural network controls a trade-off between learning speed and robustness

To understand the evolution of neural codes that might support context-dependent decision-making, we trained neural networks with gradient descent on two tasks that involved predicting the relevant feature value of a stimulus that varied systematically along two different orthogonal dimensions, of which only one was relevant for each task. Input images contained Gaussian "blobs" whose mean along the x and y coordinates were varied in five discrete steps respectively. The networks were trained in two interleaved contexts (task A and task B), signalled to the network via unique input nodes, to report either the mean x- or y-location, depending on whether the trial belonged to context A or B (**Fig. 3.1A,D**). As expected from theoretical results (Chizat et al., 2020; Woodworth et al., 2020), the norm of the weights at convergence (**Fig. 3.1B**, upper) and overall change in input-to-hidden layer weights over learning (**Fig. 3.1B**, lower) depended strongly on initial connection strengths (*Kruskal-wallis test on* $w_{hidden} : H = 235.269, p < 0.0001$*, Kruskal-wallis test on* $\Delta w_{hidden} : H = 234.51, p < 0.0001$), while the change of readout weights was substantial and differed only slightly across training regimes (*Kruskal-wallis test on* $w_{out} : H = 199.5, p < 0.0001$*, Kruskal-wallis test on* $\Delta w_{out} : H = 172.077, p < 0.0001$, **Fig. 3.1C**). Hereafter, we refer to the extremes along the continuum of weight variances as rich ($\sigma = 0.01$) and lazy ($\sigma = 3.0$) regimes. Applying PCA to the hidden layer patterns revealed that the final

representations were lower dimensional under rich learning, with just 6 (9) principal components needed to explain 95% of the variance under rich (lazy) learning (**Fig. 3.1E**). Critically, however, the rich regime proved more tolerant to a challenge that reduced the dimensionality of hidden unit activity: only 3/6 components were needed to maintain ceiling performance, whereas 8/9 were required under lazy learning (**Fig. 3.1F**). Although learning was up to 10 times faster in the lazy regime (*convergence speed lazy > rich:* $T(29) = 125.846, p < 0.0001$, **Fig. 3.1G**), the highly structured representations acquired during rich learning conferred robustness, also making performance more tolerant to the addition of Gaussian input noise (*Accuracy rich > lazy:* $T(29) = 14.55, p < 0.0001$, **Fig. 3.1H**). In other words, networks initialised in the lazy regime rapidly learned to solve the task by reading out from an approximately fixed non-linear high-dimensional random representation, whereas those initialised in the rich regime converged more slowly but exhibited strong representation learning in the input-to-hidden weights. These solutions offer complementary costs and benefits for representation learning (speed vs. robustness) of task-related variables.

3.2.2 Neural network simulations suggest two possible representational schemes for context-dependent decision making

Next, we used representational similarity analysis (RSA) and multidimensional scaling (MDS) to visualise the neural geometry of the network hidden units at convergence under either regime. During lazy learning ($\sigma = 3.0$) the similarity is mostly driven by the structure of the input space (including the task context) (**Fig. 3.2A**); this is expected because the input weights remain close to their initial values and random high-dimensional projections approximately preserve distances between inputs (Gao et al., 2019). However, during rich learning ($\sigma = 0.01$) hidden unit activity varies with context: in task A, neurons code for dimension *x* but not *y*, with the converse true for task B. In other words, task-irrelevant features were attenuated in each context, transforming the neural "grid" into two manifolds, each coding for a task-relevant dimension. Specifically, each context has a compressed and uncompressed axis, forming a rectangle in the plane, and we hereafter call the geometry "orthogonal" when the respective compressed and uncompressed axes are perpendicular across tasks. Thus, the network

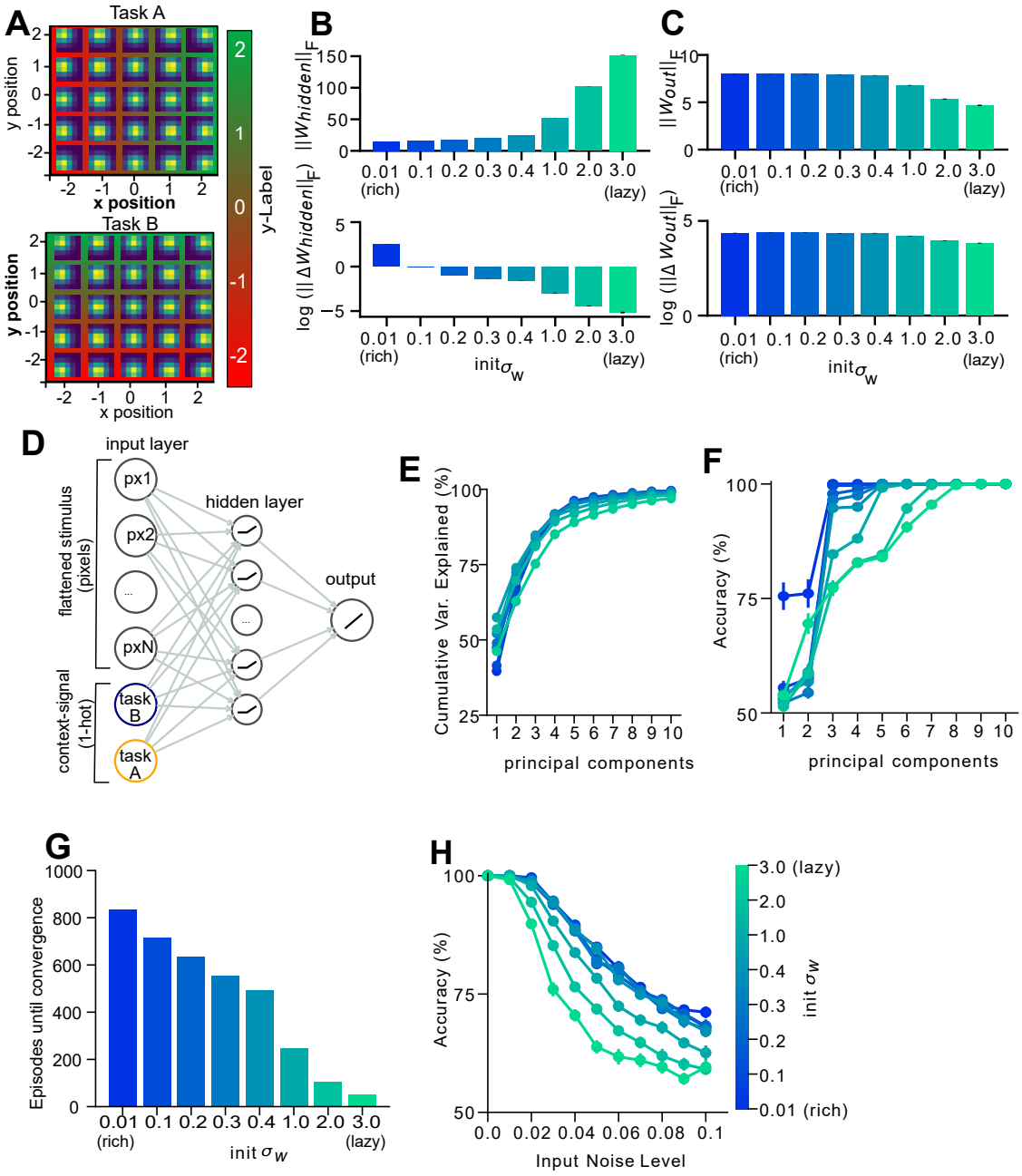

Figure 3.1: Neural Network architecture and effect of weight scale on learning speed and robustness. (**A**) In each context, the network had to predict either the x-or y-position of the mean of two-dimensional Gaussian blobs. (**B**) Norm of the hidden weights at convergence (upper panel) and overall change in weights from input to hidden layer (lower panel), both varied with initial weight scale (x-axis and green-blue colour scale). (**C**) Same as (B), but for hidden-to-output weights. (**D**) We trained a feedforward neural network with a single hidden layer of ReLU nonlinearities on the tasks. Inputs were flattened images of gaussian blobs and a one-hot encoded context cue. (**E**) Variance explained after the retention of 1-10 principal components of hidden layer activity (x-axis) under different initial weight scales. (**F**) Network accuracy as a function of retained components. Note that the rich networks (lower initial weight scale) are more robust to compression. (**G**) Episodes to convergence as a function of initial weight scale. Lazy networks converge faster. (**H**) Network performance with differing levels of input noise. Rich networks are more resilient to noise.

learned to transform the inputs in a way that might minimise intrusions from irrelevant features in each context (**Fig. 3.2B**)(Koay et al., 2019). This was confirmed by fitting model representational dissimilarity matrices (RDMs) to the hidden unit patterns at convergence (**Appendix A, Fig. A.1**): a grid model that encoded the space spanned by the two feature dimensions and context fit best for lazy solutions and an orthogonal model that encoded only task-relevant dimensions along orthogonal axes fit best for rich solutions (*grid model, lazy > rich:* $T(29) = 29.02, p < 0.0001$*; orthogonal model, rich > lazy:* $T(29) = 20.26, p < 0.0001$, **Fig. 3.2C**) Both models explained the patterns better than a parallel model that represented an encoding of the value of stimuli along two parallel planes, obtained by rotating one of the task manifolds from the orthogonal model by 90 degrees (*grid model > parallel model:* $T(29) = 74.69, p < 0.0001$*, orthogonal model > parallel model:* $T(29) = 82.61, p < 0.0001$, **Fig. 3.2C**). We also used RSA in conjunction with a parametric model-fitting approach. Rather than fitting models encoding extremes of compression, rotation, and context separation, now we built RDMs by varying these factors continuously. Fitting this parameterised model to the neural network data confirmed that compression along irrelevant dimensions was larger under rich than lazy learning ($T(29) = 49.77, p < 0.0001$, **Fig. 3.2D**). The estimated rotation parameter was close to zero (**Fig. 3.2E**) which suggests that information was kept in the frame of reference of the inputs, yielding orthogonal and grid-like representations in the rich and lazy regime respectively. In both regimes, a third dimension encoded context, indicated by a non-zero offset parameter (**Fig. 3.2F**). Taken together, a simple neural network can solve the tasks either by employing high-dimensional and task-agnostic or low-dimensional and task-specific representations. The variance of weights at initialisation determines how learning dynamics shape representational geometry.

3.2.3 Task specific representations can be achieved via non-linear gating

How does this neural coding scheme prevent interference among tasks? Early work on attention and cognitive control suggests that a gating mechanism could be employed to selectively activate units that encode information that is relevant for the task (Miller

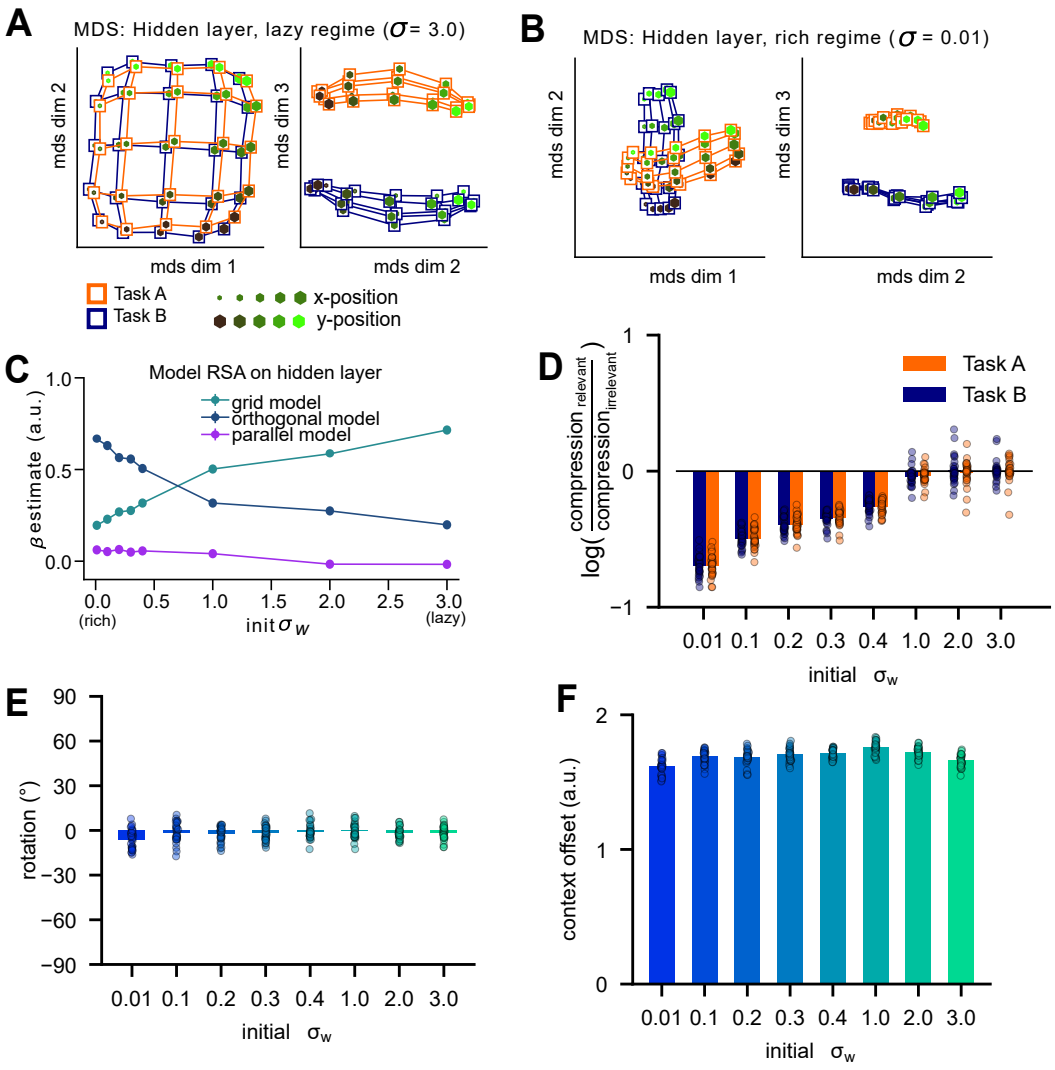

Figure 3.2: Geometry of representations in hidden layer of trained neural network. (**A**) 3D representation of hidden layer representations for each stimulus feature (x- and y-position, dot colour and size) in each context (connecting lines, orange and blue) after training in the lazy regime. (**B**) Same as (A) but for training in rich regime. Note the emergence of orthogonal manifolds which compress along the irrelevant dimension aligned with dimensions 1 and 2. (**C**) Fits of RDMs encoding grid, orthogonal and parallel representational schemes to the neural network data as a function of initial weight scale. The orthogonal model (dark blue line) fits best in the rich regime, and the grid model (cyan line) fits best in the lazy regime.(**D-F**) Estimates for compression, rotation and offset of the parameterised RSA model. Best-fitting RDM characterised by parametrically varying expansion/contraction of representation on relevant/irrelevant dimension (D), context-dependent rotation of the stimulus axes from native space into the reference frame of the response (i.e. from orthogonal to parallel model, E) and separation between contexts (F).

& Cohen, 2001). The classic model of cognitive control implements this gating with hard-coded biases that move activity in and out of the linear range of sigmoidal non-linearities in the network's hidden layer (Cohen et al., 1990). In contrast to this earlier work, our neural network is trained end-to-end on the task without enforcing the gating scheme by hand. How then, do task-specific representations emerge under rich learning? We reasoned that orthogonal manifolds could emerge if the weights linking each context unit to the hidden layer were anticorrelated. Sufficiently large anticorrelated weights ensure that distinct subsets of hidden units are active in each context, as neurons which receive negative net input in one context (and which therefore are inactive due to the rectified linear (ReLU) activation function) will receive positive net input (and be active) in the other. By wiring only the task-relevant stimulus dimension to the active population in each context, information along the irrelevant dimension is thus effectively zeroed out by the nonlinearity, creating an independent subspace for each task (**Fig. 3.3A**). This would allow the network to factorise the problem, encoding the task-relevant information in a way that avoids mutual interference (**Fig. 3.3B-D**).

3.2.4 Empirical evidence in neural networks supports gating theory

This theory makes several testable predictions. Firstly, it implies that most neurons should be mixed selective, responding to combinations of stimuli and task variables. Secondly, however, it implies that this mixed selectivity should be structured in the rich regime, with most units in the hidden layer responding specifically to the combination of task-relevant stimulus dimension and task. Indeed, we observed that up to 60% of hidden units responded exclusively under one task or the other during rich learning (**Fig. 3.4A**). Visualising the receptive fields of the hidden layer units revealed that under rich learning, weights into task-specific units were aligned with the relevant feature dimensions, whereas under lazy learning, more heterogenous selectivity patterns were observed (**Fig. 3.4B,C**). Investigating the magnitude of the readout weights confirmed that the rich-initialised network relied mostly on those task-specific units to make predictions (**Fig. 3.4D**). Thirdly, the theory predicts that in neural networks the context weights should be anti-correlated. This is indeed the case on average in

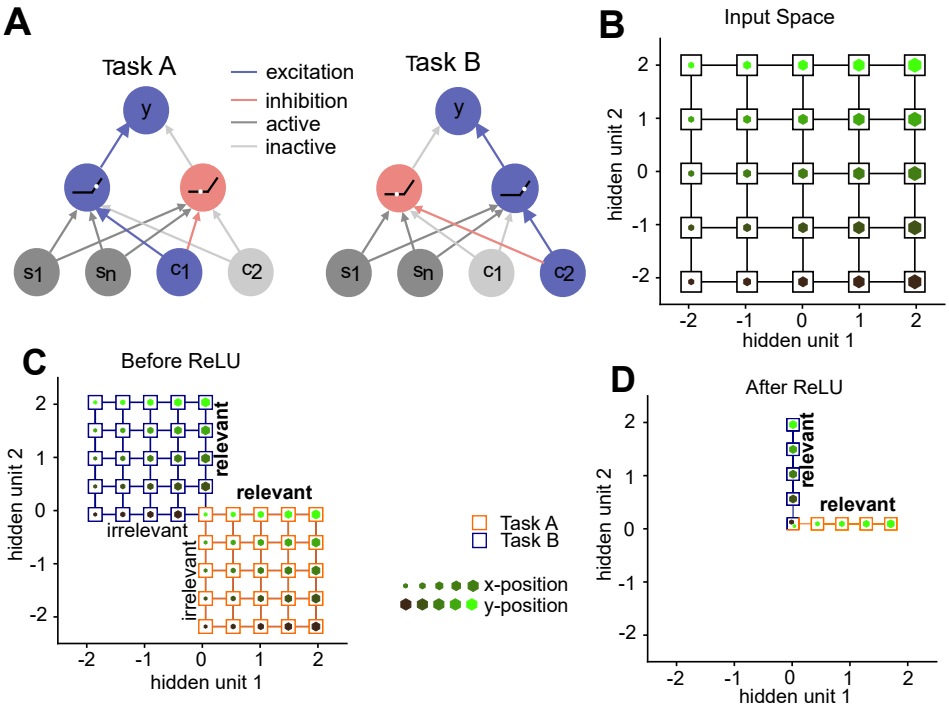

Figure 3.3: Nonlinear gating theory. (**A**) Schematic illustration of how opposing weights from two context units leads to learning two unique subspaces. Red and blue arrows show positive and negative weights from context units, which control the sign of the net inputs in the hidden layer, so that stimuli are effectively processed by different hidden units in each context. (**B-D**) Schematic illustration but in neural state space. (B) Shows similarity structure among input stimuli with no context modulation. (**C**) Shows the similarity structure in the hidden layer net input (before ReLU). Note the separation between contexts. (**D**) After the ReLU, "inhibited" (below-zero) inputs are removed, leaving two orthogonal manifolds.

the rich regime (**Fig. 3.4F**) and especially for most task-specific neurons (**Fig. 3.4G**), which became anti-correlated as training progressed. In contrast, those neurons that converged to being task-agnostic were those that received strong, positively correlated input from two context units at random initialisation, and this input remained positively correlated after training (**Fig. 3.4H**). It thus seems likely that the initial sign of the connections from the context units to each hidden unit determines whether it is destined to be a task-agnostic or task-specific unit during training. Furthermore, the theory predicts that task-specific units show a coding preference for relevant feature dimensions (with irrelevant features mapped onto units which are deactivated by the ReLU). This is exactly what is seen in the neural network, where the responses of task-

specific units are aligned to the two choice axes (**Fig. 3.4I-J** *factorised model > linear model:* $z = 4.781, p < 0.0001, d = 0.873$). The remaining 35% of active units coded for a residual policy which collapses across both contexts ("task agnostic"), resembling the linear model described above (**Fig. 3.4I-J** *linear model > factorised model* $z = 4.781, p < 0.0001, d = 0.873$).

A final prediction of this theory is that in the rich regime, performance depends critically on the task-specific neurons but not on those displaying task-agnostic selectivity. We thus conducted an ablation study in which the output of either the task-agnostic or task-specific neurons was set to zero at evaluation. How did performance depend on these units? The analyses of the receptive fields suggest that under rich learning - but not under lazy learning, removing task-selective units should only impair performance on one task, but not on the other task for which units were not selective. This is in fact what we observed (**Fig. 3.5A**). As task-agnostic units ignored the context under rich learning, removing the task-specific units should impair performance on incongruent, but not on congruent trials. Again, this was confirmed by our simulations (**Fig. 3.5B**). Together, these findings support a model of context-dependent decision-making whereby the network learns to gate information into orthogonal subsets of hidden units in a way that minimised mutual interference. This scheme emerges when context input weights are anticorrelated.

3.2.5 Parallel representations can emerge in deeper neural networks

We observed that rich learning induced a task-specific geometry in the hidden layer, with different subsets of units coding for different tasks. At the population level, this led to an orthogonal representation, coding for the context and the relevant dimensions of each task along separate axes. Our gating theory can explain how these patterns emerge. However, in both contexts, the network had to learn the signed distance of a stimulus to the category boundary. Hence, it could have learned a general representation of response magnitude, in which features lie on parallel axes in coding space. Indeed, previous work suggested that generalisable task features lie on parallel planes (Bernardi et al., 2020). In the previous set of simulations, we tested this hypothesis

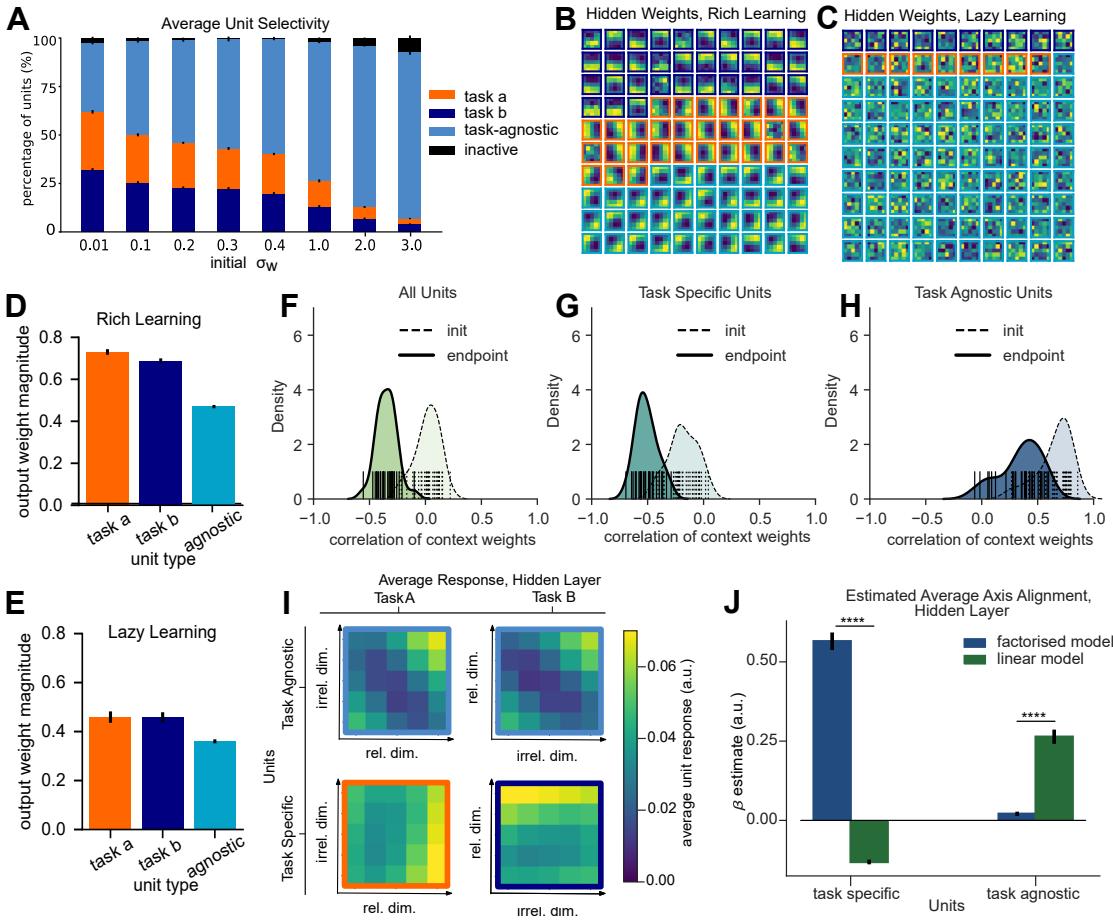

Figure 3.4: Neural network data in support of gating theory. (**A**) Proportion of task-agnostic and task-specific units in the neural network as a function of initial weight scale. (**B**) Hidden layer weights after training in the rich regime. Each subplot corresponds to the weights from the input layer to a single hidden unit, reshaped to the original dimensions of the input. Frames around subplots indicate task-selectivity as defined in (A). (**C**) Same as (B) but for lazy regime. (**D**) Magnitude of readout weights after rich learning, shown separate for weights originating in task-a, task-b selective, or task-agnostic units. (**E**) Same as (D) but after lazy training. (**F**) Distribution of empirically observed correlation coefficients among context unit weight vectors in the neural network. (**G-H**) Same as (F) but separated out by "task-specific" and "task-agnostic" units as defined in (A). Note the anticorrelation in task-specific units (and overall). (**I**) Hidden unit selectivity for each relevant and irrelevant stimulus feature in each context. Note that task-specific units (lower panels) are mostly sensitive to relevant vs. irrelevant dimension whereas task-agnostic units code for an interaction between features. (**J**) Quantification of results in (I) using fits of linear vs. factorised model. The factorised model fits best to task-specific units, and the linear model to task-agnostic units.

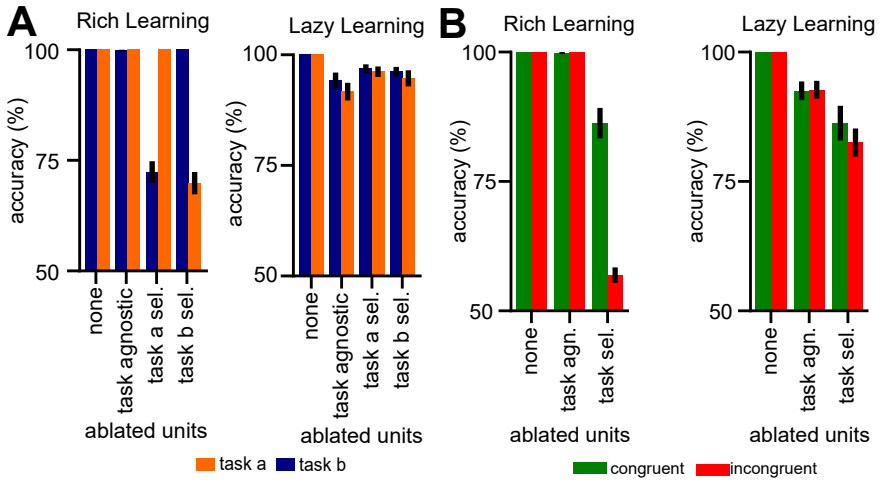

Figure 3.5: Results of ablation study. (**A**) Ablation study, with performance shown separately for task A and B. Under rich learning, where task-specific units are axis aligned, removing them only affects performance on the task they are selective for. (**B**) Ablation study with performance separately on congruent (same response in both contexts) and incongruent (different responses in task A and B) trials. Under rich learning, task-specific units are axis aligned, and task-agnostic units encode congruent trials. Hence, when task-specific units are removed, the network still performs well on congruent, but not incongruent trials.

explicitly with the parallel model RDM, but the orthogonal model explained the hidden layer geometries much better. To test whether a network with a single hidden layer could in principle acquire a parallel representation of the task-relevant axes, we added a regularisation term to the loss function that quantified the mismatch between an RDM computed from the hidden layer activity patterns and a target RDM that encoded parallel representations (**Fig. 3.6A**, methods). We then performed a random hyperparameter search to find values that would allow the network to learn the tasks with a representation constrained by this additional loss. As control, we trained the same model without this "RDM loss", as well as variants with grid-like and orthogonal target RDMs. All models converged to ceiling training performance (**Fig. 3.6B**). However, while this procedure was successful in biasing the network towards either grid-like or orthogonal representations when the respective RDMs was used as target, it failed to acquire a parallel representation (**Fig. 3.6C-D**. *parameter estimates, parallel RDM-loss: orthogonal* $>$ *grid:* $z = 3.9125, p < 0.0001$*; orthogonal* $>$ *parallel:* $z = 3.9125, p < 3.9125$). We speculated that this might be different for deeper networks, where task-irrelevant

information could first be filtered out, before the representation would be transformed into the reference frame of the response. To test this, we repeated the simulations in a network with two hidden layers and applied the RDM loss to representations in its second hidden layer (**Fig. 3.7A**). Again, the network converged to ceiling performance on both tasks under interleaved training (**Fig. 3.7B**). This time, however, it was also able to learn a parallel representation in its second layer, when pressured to do so with a parallel target RDM (**Fig. 3.7C-D**; *parameter estimates, parallel RDM-loss: parallel* $>$ *grid:* $z = 3.9125, p < 0.0001$*; parallel* $>$ *orthogonal:* $z = 3.9125, p < 3.9125$). Interestingly, we observed orthogonal patterns in the first hidden layer, consistent with our predictions (**Fig. 3.7D**).

3.2.6 Task-specific versus task-agnostic representations under rich and lazy learning in CNNs

In the introduction of this thesis, I argued that simple neural network models are useful to generate predictions for the computations that might underlie complex cognitive processes. To test whether our findings would generalise to complex architectures, we trained deep convolutional neural networks directly on images from a task that had been previously used to study continual learning in humans (Flesch et al., 2018) and investigated the representations formed in the hidden layers under rich and lazy learning (**Fig. 3.8A-B**, methods). Stimuli were fractal images of trees that varied in their density of leaves (leafiness) and branches (branchiness) and pasted on a grey background, surrounded by either an orange or blue rectangle to indicate whether they belonged to context A or B. In each context, only a single dimension, either branchiness or leafiness, was relevant. Like the Gaussian "blobs", these dimensions were varied parametrically in five discrete steps, spanning a 5x5 grid of possible feature combinations, and in each context (task) the network had to predict the signed distance to the category boundary along a single task-relevant dimension (either leafiness or branchiness, depending on context) (**Fig. 3.8A**, see methods for details). Irrespective of the learning regime, the networks converged to about 80% validation accuracy on a held-out test set (**Fig. 3.8C**). While we did not observe significant differences in generalisation performance or convergence speed on this dataset, rich learning induced highly struc-

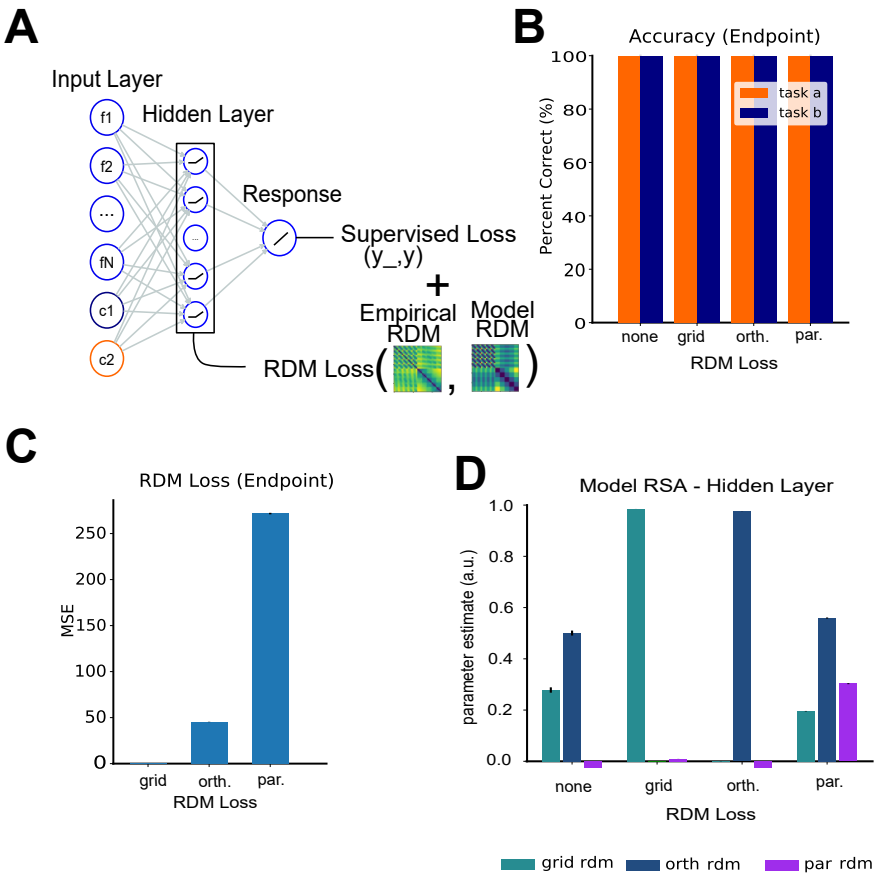

Figure 3.6: Results of simulations with auxiliar RDM loss. (**A**) We equipped the network with an auxiliary objective ("RDM loss") which minimised the difference between patterns in the hidden layer and a candidate model RDM that encoded either grid-like, orthogonal or parallel representational schemes. (**B**) Accuracy after convergence on the supervised objective, depending on chosen constrain on RDM. All models converged. (**C**) Endpoint RDM loss after convergence on the supervised objective. All networks except for the one with parallel model RDM loss converged. (**D**) Model RSA. The models with grid and orthogonal schemes as target for the RDM loss learned the desired representations. The model trained with a parallel RDM as target in the RDM loss converged to orthogonal representations.

tured representations that progressively transformed the inputs from more grid-like representations in the early layer, over orthogonal representations in the intermediate layer to parallel representations in the deep layers (**Fig. 3.8D**). In contrast, under lazy learning, all hidden layers exhibited task-agnostic representations (**Fig. 3.8E**). Just as for the simpler MLPs, we also fitted a more flexible parameterised model RDM to the representations in each hidden layer, finding that under rich learning, irrelevant dimen-

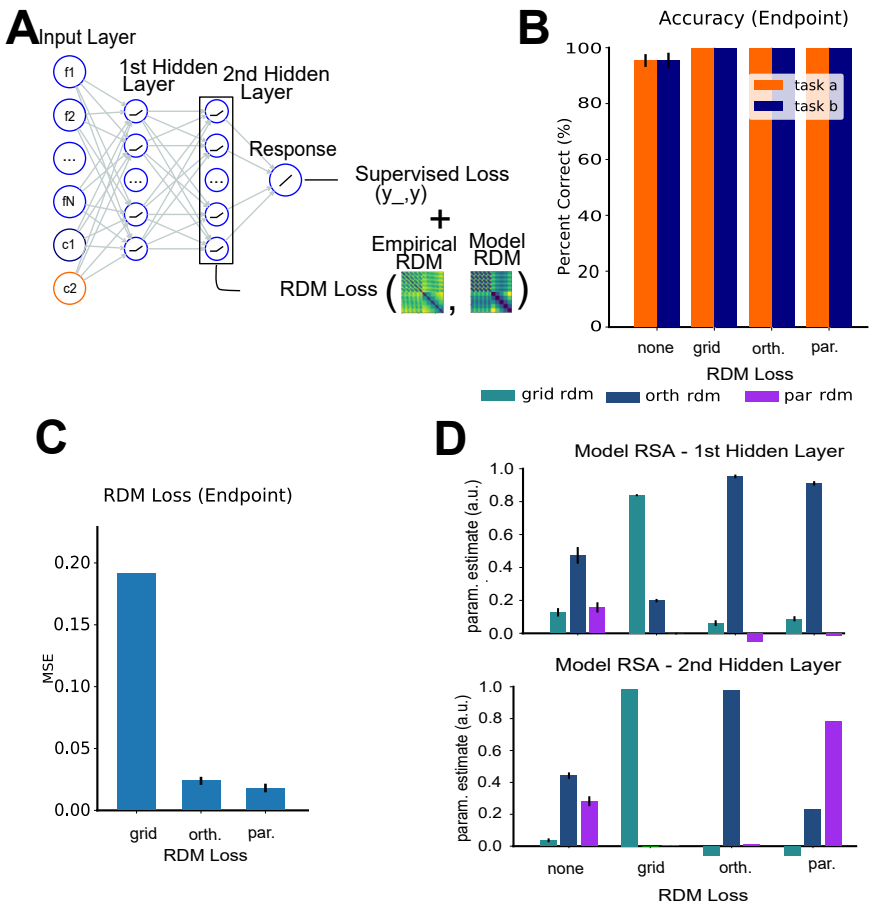

Figure 3.7: Results of simulations with auxiliar RDM loss., 2-layer MLP (**A-D**) Same as (Fig. 3.6A-D) but for model with two hidden layers. This time, parallel representations could be enforced in the second layer, leading to orthogonal representations in the first layer.

sions were suppressed more in deeper layers, and that representations of the relevant features were eventually rotated into the same reference frame (**Fig. 3.9A**). In contrast, representations under lazy learning did not change substantially across the hierarchy of the neural network (**Fig. 3.9B**).

3.3 Discussion

The work described in this chapter makes several distinct contributions. The first is to formalise solutions to the learning of a canonical context-dependent classification paradigm using a feedforward connectionist (or "deep learning") framework (Richards et al., 2019; Saxe et al., 2021). We do this by drawing upon recent work in machine learning research, which distinguishes among the learning regimes which occur when

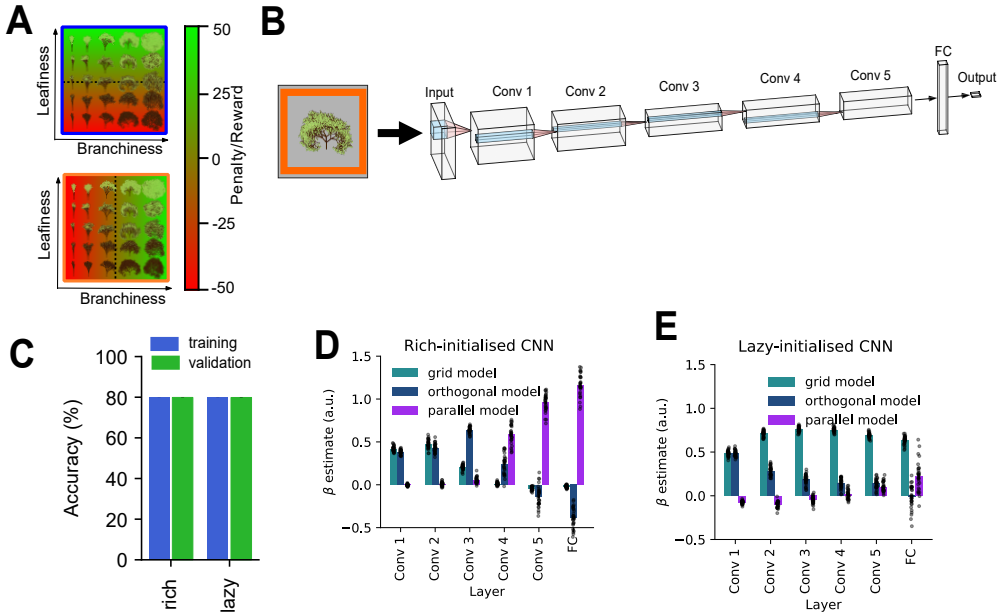

Figure 3.8: Convolutional Neural Network (CNN) trained on high-dimensional stimuli. (**A**) Stimulus and task-design, adapted from (Flesch et al., 2018). Each image shows the category boundary (dashed line) and reward/penalty (red-green colour) for choosing to "plant" a tree in a specific context (signalled by blue frame/orange frame). In the original study, participants were asked to "accept" (plant) or reject a tree by pressing one of two buttons. Frame colour signalled context. Here we trained the network to predict this reward associated with each stimulus. (**B**) Network architecture. We trained a feedforward CNN with five convolutional and one full-connected layer on the trees task. The network received RGB images of trees surrounded by an orange/blue frame to signal context, and had to predict the task-relevant label (level of branchiness/leafiness). The network was trained either in the rich (small initial weights) or lazy (large initial weights) regime. (**C**) Accuracy after convergence, separately shown for training and validation dataset. No difference in performance between training regimes. (**D**) RSA results for network trained in rich regime. Coefficients of grid, orthogonal and parallel model obtained from Linear Regression performed on patterns from each layer. Each dot corresponds to a single trained neural network (30 in total). Early layers encode both feature dimensions, followed by orthogonal representations in intermediate layers and parallel representations closer to the readout. (**E**) Same as (D), but for network trained in lazy regime. Little evidence for representational change in the network.

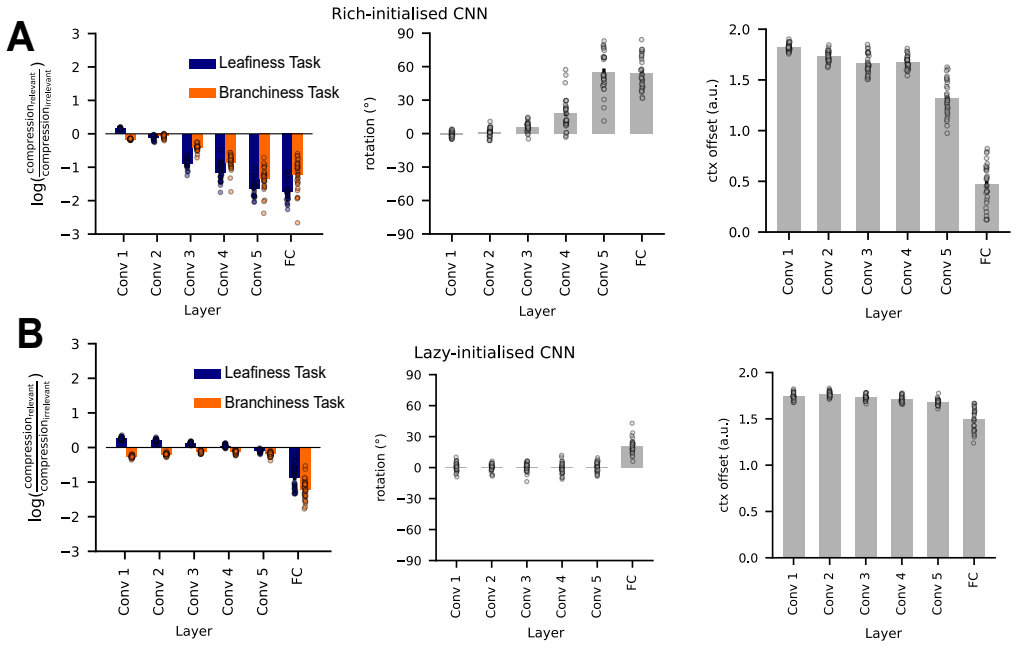

Figure 3.9: Parametrised model, fit to representations in CNN. (**A**) The conversion from task-agnostic representations of the inputs into task-specific representations of the rules was confirmed by fitting the compression, rotation and offset parameters of the fully-parameterised model RDM. (**B**) Same as (A), but for model trained in the lazy regime. All convolutional layers had task-agnostic representations, while the FC layer showed signs of task-specific representations.

deep networks are initialised with high variance (lazy) or low variance (rich) weights (Arora et al., 2019; Chizat et al., 2020; Fusi et al., 2016; Geiger et al., 2020; Jacot et al., 2020; Lee et al., 2019; Rigotti et al., 2013; Woodworth et al., 2020). We derive predictions from these regimes for the context-dependent decision-making task, a paradigm that has been well-studied before using both single neuron electrophysiology (Aoi et al., 2019; Mante et al., 2013) and neuroimaging (Takagi et al., 2020) methods. The second contribution is an insight into the computational principles that allow the context-dependent decision task to be solved. We show that a combination of anticorrelated context inputs and ReLU (or ReLU-like) non-linearities allows the network to effectively learn to gate task information according to context. This allows us to predict how mixed-selective units code for relevant and irrelevant features in the neural network, and to anticipate the effects of silencing task-agnostic vs. task-specific neurons on performance. The third contribution are insights into the impact of network depth on the type of representations that emerge under rich learning. Both, through ex-

plicitly pressuring the network to learn a certain representation, and by training a much more complex network on naturalistic stimuli, we provide evidence for several subsequent processing steps which progressively transform inputs from a task-agnostic into a highly task-specific format. More specifically, we observed orthogonal representations in intermediate layers, whereas layers close to the output encoded information in the response frame of reference, with relevant information of both tasks aligned along parallel planes.

There has been a recent resurgence of interest in neural networks (or "deep learning models") as computational theories of biological brains (Richards et al., 2019; Saxe et al., 2021). A common approach is to use methods such as RSA to examine similarities between the representations formed in biological systems (e.g. multi-neuronal or multivoxel patterns) and in the hidden units of deep networks. One corollary of our findings is that the relationship between representations formed in biological and artificial networks can critically depend on the variance of the weights at initialisation. For example, when the initial weight scale is large, the similarity structure of encoded representations will closely match their input structure. This may partly explain why previously reported improvements in model fit of trained over untrained networks tend to be relatively modest, as if the visual cortex mainly recapitulates the input data through random high-dimensional projections (Guclu & van Gerven, 2015; Schrimpf et al., 2018).

We manipulated whether a network operated in the rich or lazy regime by changing the scale of the weights at initialisation (Chizat et al., 2020). One interpretation of this result is that whether a circuit operates in one regime or the other might depend on its initial connection strength, and hence the prior information that is encoded in this network. The networks trained in the rich and lazy regime converged at different rates, suggesting that the chosen learning rate might have an impact on the representations acquired by the networks. However, repeating the simulations for a range of different learning rates revealed that while this hyperparameter choice had some impact on the weight change, the overall difference in weight changes and representations between rich and lazy learning was independent of this manipulation (**Appendix A, Methods**

and Figure A.2). We note, however, that the same results can be obtained by adding an L2 regulariser to a network initialised in the lazy regime (**Appendix A, Methods**). Increasing the regularisation strength has the same effect as initialising the network with smaller weights (**Appendix A, Figure A.3**). From a mathematical perspective, this equivalence may not be surprising, as the L2 regularisation can be seen as putting a Gaussian prior with variance given by the inverse of the regularisation strength on the distribution of network weights (Hastie et al., 2009). These complementary simulations predict that the brain could potentially change between the two regimes in a flexible manner, depending on the specific task demands, a hypothesis which remains to be tested.

We show that different representational motifs can result in similar task performance, but that these might trade-off learning speed for robustness. This furnishes precise predictions for the type of neural geometry one would expect to find in recordings from human or non-human primate brains, and the specific error patterns one would observe in behaviour. Our simulations with deeper networks suggest that if the brain operated in a rich regime, different representational geometries would to be expected in different areas. Earlier work revealed a striking correspondence between EVC and early convolutional layers (Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). Our model trained in the rich regime would predict that representations in EVC are mostly task-agnostic. Gating theories of cognitive control have singled out lateral portions of the frontal lobe as locus for task-specific processing (Miller & Cohen, 2001b). In line with our gating theory of rich learning, which builds directly on these theories of cognitive control (Cohen et al., 1990), we might expect to find orthogonal representations in these prefrontal regions. Parallel representations would then be expected in regions that prepare and execute the motor response.

These insights hint at the possibility that the precise geometry of representations learned under rich learning might also depend on the specific task demands. Here, we investigated a paradigm in which different features of the same type of stimuli were relevant in each context, with rules being orthogonal by design in this feature space. An influential study that reported parallel planes in prefrontal regions used a different

paradigm with fractal images that had no parametrically varying perceptual features (Bernardi et al., 2020). Hence, it is likely that specific task demands, and the extent to which different features can be shared across tasks or need to be separated to avoid interference, influence the geometry of the acquired neural representation.

It should be noted that all networks were trained on interleaved data, rather than sequentially on one task after the other, to prevent catastrophic forms of forgetting (French, 1999). Future work could investigate whether one learning regime might be less prone to forgetting than the other, and which additional computational mechanisms would be required to allow these networks to have learning dynamics similar to those observed in humans (Flesch et al., 2018).

Taken together, this work provides a computational framework to study the type of representational geometries that might develop when learning context-dependent decision-making problems and concrete testable predictions regarding the format and location of these representations in brains.

3.4 Methods

3.4.1 Neural network simulations with varying weight scale

The simulations were implemented, and results analysed in Python using the NumPy, SciPy and Scikit-Learn packages. Due to the simplicity of the architecture, gradients, and optimisation procedures for the simple feedforward MLPs were derived by hand and implemented in raw NumPy. The more complex simulations with auxiliary RDM loss were implemented in Tensorflow.

Task design: Stimuli were images of two-dimensional isotropic Gaussian "blobs". The stimulus space was spanned by parametric modulation of the x and y coordinates of these blobs in five discrete steps. Inside this 5x5 grid, neighbouring blobs were partially overlapping, allowing the networks to infer similarity structure based on co-activation of input units. The networks were trained on a context-dependent decision-making problem. There were two contexts, in each of which only one feature dimension (either the x- or y-location) was diagnostic of the correct output (the other being an irrelevant dimension) and mapped onto a numerical value ranging from -2

to 2, which denoted the signed distance to the category boundary along the relevant feature dimension. The network was trained to predict the relevant feature value associated with each stimulus. To assess performance and representational geometries, we fed trials covering all combinations of the two feature dimensions (x/y location) and context into the network and recorded hidden layer activity patterns as well as network outputs for each stimulus.

Neural network architecture: Our model was a feed-forward network architecture with a single hidden layer. Input units encoded pixel intensities of vectorised and normalised images of Gaussian blobs. Each image had a down-sampled resolution of 5x5 pixels, resulting in 25 stimulus input units. Two additional one-hot encoded inputs (1 or 0) signalled the context to the network. All 27 inputs were projected into a hidden layer with 100 units and Rectified Linear Unit (ReLU) nonlinearities. The responses of the hidden units were mapped onto a single linear output unit.

Weight initialisation: All network parameters were initialised with random draws from Gaussian distributions with a mean of zero. To control whether the network operated in the rich or lazy regime, we modified the variance of these distributions systematically, ranging from $\sigma = 0.01$(rich regime) to $\sigma = 3$ (lazy regime). We call this "initial weight scale" in the main text. These values were derived empirically by observing their impact on the relative change of the weight norm and shape of the loss trajectories during training. Weights to the output unit were instead initialised with a variance scale of $\frac{1}{\sqrt{n_h}}$ where n_h is the number of hidden units. All biases were initialised to zero.

Training: We collected 30 independent runs (unique random initialisations) per initial weight scale. On each run, the network was trained with minibatch gradient descent (batch size 50, interleaved data, learning rate $\varepsilon = 0.001$, SGD optimiser) on 10000 iterations. The model was trained on the Mean-Squared-Error (MSE-Loss) between the true and predicted relevant feature associated with each stimulus:

$$J(W) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i, W))^2 \tag{3.1}$$

Addition of Gaussian input noise: We investigated the robustness of different training regimes to additive Gaussian noise in the inputs. The model architecture and training procedures were identical to the ones described above. Again, we collected 30 independent runs per weight scale, ranging from 0.01 to 3 in eight steps. However, this time, we added Gaussian noise drawn from a standard normal distribution to the input units at test. The strength of this noise was varied parametrically in 10 steps from $\sigma = 0$ to $\sigma = 0.1$, allowing us to investigate the impact of different noise levels on performance.

3.4.2 Neural network simulations with auxiliar RDM loss

To test what kind of representations a network with a single hidden layer could theoretically acquire when trained in the rich regime, we carried out a new set of simulations in which we introduced an auxiliary loss function that quantified the mismatch between an RDM constructed from the hidden layer patterns and a model RDM that encoded a target representational geometry and repeated the simulations described above. On each training step, a minibatch of all 50 stimulus types was passed into the network. In the hidden layer, this yielded a 100x50 activity matrix Y_{hidden}. We computed a 50x50 RDM from these activity patterns (Euclidean distance) as follows:

$$G = {Y_{hidden}}^T Y_{hidden} \tag{3.2}$$

$$RDM_{hidden} = diag(G) + diag(G)^T - 2G \tag{3.3}$$

We then calculated the mean squared error between this RDM and a target RDM, which was chosen to be either the grid, orthogonal or parallel model RDM (see section on Representational Similarity Analysis in methods for details of those models). The total loss of the network was a weighted sum of the standard supervised objective (the MSE between the network's output and the target label) and this RDM loss:

$$J(W) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, W))^2 + \frac{\beta}{2}\left(vec(RDM_{target}) - vec(RDM_{hidden})\right)^2 \tag{3.4}$$

This encouraged the network to learn the task whilst being pressured to acquire the representation imposed by the RDM loss. We performed a random search to find hyperparameters that allowed the network to learn the task with grid, orthogonal or parallel representations in the hidden layer. However, as the model was unable to acquire a parallel representation, we then introduced a second hidden layer (100 ReLUs) to the network to increase its capacity and repeated the procedure with the RDM loss applied to the second layer. Again, we collected 30 independent training runs per simulation.

3.4.3 Convolutional neural network simulations

To test whether the findings reported with a small MLP generalise to more realistic settings, we trained a deep convolutional neural network (a variant of AlexNet) on context-dependent decision-making problem with fractal tree images. The network was implemented in PyTorch.

Task design: Stimuli were procedurally generated images of trees for which we varied the density of leaves (leafiness) and branches (branchiness) in five discrete steps. There were two contexts, signalled by an orange/blue frame around the image of a tree. The entire dataset consisted of 20000 training and 10000 test RGB images (96x96x3 pixels) of fractal trees, pasted on a grey background and surrounded by an orange rectangle in context A, or blue rectangle in context B. In each context, only a single dimension (either leafiness or branchiness) was relevant and mapped onto the same numerical target signal as described for the neural network simulations above.

Neural network architecture: The CNN was a variant of AlexNet and consists of a series of five convolutional layers with/without max-pooling (1st Conv2d: 64 filters, size 11, stride 4, padding 2, ReLU, max-pooling size 3, stride 2; 2nd Conv2d: 192 filters, size 5, padding 2, ReLU, max-pooling size 3, stride 2; 3rd Conv2d: 384 filters, size 3, padding 1, ReLU; 4th Conv2d: 256 filters, size 3, padding 1, ReLU, 5th Conv2d: 256 filters, size 3, padding 1, ReLU) followed by a fully-connected layer (512 units, ReLU) and a single linear readout unit.

Training procedure: The CNN was trained with minibatch gradient descent (batch size 128) on the mean squared error between the network's output and the ground truth associated with each input image. Training was carried out for a total of 200 epochs.

We used an Adam Optimiser (learning rate $\varepsilon = 1e-4$). We trained the network in the lazy and rich regime by changing the variance of weights and hence their overall norm at initialisation. For each regime, we collected 30 independent training runs.

3.4.4 Quantification and statistical analysis

Accuracy: The network was trained to predict the value of the relevant feature dimension in each context, defined as the signed distance to a category boundary, $y \in [-2,-1,0,1,2]$. In contrast, in our previous behavioural study, human participants had to accept/reject stimuli based on this signed distance. For comparison between neural networks and human participants, we quantified the network's accuracy as the match between the signs of the network's predictions and the ground truth:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{n} \mathbf{1}_{((\hat{y}_i > 0) == (y_i > 0))} \tag{3.5}$$

Endpoint weight norm and relative weight change: Every 100 epochs during training, we computed the Frobenius norm of the hidden layer weights

$$||W_{hidden}||_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} |w_{ij}|^2} \tag{3.6}$$

and their relative change with respect to the norm at initialisation. This allowed us to assess whether the network operated in the rich or lazy regime, corresponding to low and high norm solutions. The weight change relative to initialisation was quantified by computing how the norm of the hidden layer weights changed from random initialisation to the endpoint of training.

$$log(||\Delta W_{hidden}||_F) = log\left(\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} |w_{ij}^{t=T} - w_{ij}^{t=0}|^2}\right) \tag{3.7}$$

Representational Similarity Analysis: We performed RSA on the hidden layer activity patterns to assess how training sculpted the representations formed by the neural network. For each individual run, we calculated RDMs based on the hidden layer activity patterns evoked by inputs covering all combinations of feature values and contexts.

The resulting 50x50 RDMs captured the Euclidean distances between all possible pairs of stimuli in the high-dimensional space spanned by the hidden units (after the ReLU nonlinearity). We visualised these geometries by projecting the group-level RDM, averaged across independent runs, down into three dimensions using metric MDS.

RSA: Quantifying hidden layer geometries: To quantify the extent to which hidden layer geometries exhibited patterns consistent with our hypotheses, we performed a linear regression of the hidden layer RDMs onto a set of model RDMs. There were three model RDMs in total, (1) a grid model, encoding the stimulus spaces as two parallel grids, separated by the context, (2) an orthogonal model, encoding the task relevant dimensions as two orthogonal 1D manifolds and (3) a parallel model, encoding the same information as the orthogonal model, but rotated into the frame of reference of the response (i.e., a "magnitude" representation). Let the vectors that denote the two feature dimensions be $b = [-2,-1,0,1,2]^T$ and $l = [-2,-1,0,1,2]^T$. Let the task vector be defined as $t = [0,1]^T$. Let the matrix of all possible ordered tuples of task, the first and the second feature dimension be:

$$X^{50x3} = \{(x,y,z) : x \in t, \;\; y \in b \;\; and \; z \in l \;\; \} \tag{3.8}$$

The first model RDM encoded two parallel, evenly spaced grids (unit distance), representing all combinations of the context and feature dimensions. This RDM was constructed by computing all pairwise Euclidean distances between the rows in X. The second model was obtained by taking the grid model and projecting stimuli onto the task-relevant axis for each context. Let X_A be the submatrix for the first task, i.e. where $t_i = 0$ and X_B the submatrix for the second task, i.e. where $t_i = 1$:

$$X_{grid} = \begin{bmatrix} X_A \\ X_B \end{bmatrix} \tag{3.9}$$

Let Y_A be the projection matrix for the first task and Y_B the projection matrix for

the second task:

$$Y_A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{3.10}$$

$$Y_B = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{3.11}$$

Then, the orthogonal model would be obtained by stacking $X_A Y_A$ and $X_B Y_B$:

$$X_{orth} = \begin{bmatrix} X_A Y_A \\ X_B Y_B \end{bmatrix} \tag{3.12}$$

Again, the corresponding RDM was obtained from this coordinate matrix by computing the pairwise Euclidean distances between its rows. Thus, for each context, stimuli differed along the task-relevant dimension (unit distance), and representations of different tasks were orthogonal to each other. The parallel model was obtained by rotating one of the task vectors from the second model by 90 degrees, so that it only encoded the signed distance to the category boundary along the respective task-relevant dimension. The rotation matrix was defined as:

$$R_A(90) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(90) & sin(90) \\ 0 & -sin(90) & cos(90) \end{bmatrix} \tag{3.13}$$

The parallel model would be obtained by stacking $X_A Y_A R_A$ and $X_B Y_B$ and computing pairwise Euclidean distances between its rows:

$$X_{par} = \begin{bmatrix} X_A Y_A R_A \\ X_B Y_B \end{bmatrix} \tag{3.14}$$

The lower triangular form of these models was z-scored and entered into a linear multiple regression model to predict the lower triangular form of the hidden layer

RDM:

$$RDM_{brain} = \beta_0 + \beta_1 RDM_{grid} + \beta_2 RDM_{orth} + \beta_3 RDM_{parallel} \tag{3.15}$$

This procedure was repeated for each individual training run, yielding a distribution of regression coefficients that permitted statistical inference on the relative difference between predictors as well as their difference from zero. We tested whether two models differed in their extent to which they covaried with the hidden layer RDM by performing Wilcoxon Signed Rank tests on their corresponding beta estimates. A non-parametric test was chosen due to the observed violation of the normality assumption. We applied this analysis to models with different initial weight scale, enabling us to investigate the impact of the training regime (rich or lazy) on the emerging representations.

RSA: Parameterised model: In order to obtain more fine-grained estimates of the learned geometry, we also fit a parameterised model to the RDMs constructed from the hidden layer patterns. We constructed a space of model RDMs by varying six parameters, one controlling the angle between the task-specific grids (which rotated one task manifold by up to 90 degrees in either direction, so that representations could be orthogonal, parallel, antiparallel or anything in between), four controlling for the compression of relevant and irrelevant dimensions within each context, and one controlling for the separation of contexts. Let the vectors denoting the two feature dimensions be $b = [-2,-1,0,1,2]^T$ and $l = [-2,-1,0,1,2]^T$. Let the task vector be defined as $t = [0,1]^T$. Let the matrix of all possible ordered tuples of task and the two feature dimensions be:

$$X^{50x3} = \{(x,y,z)\ x \in t,\ \ y \in b \ \ and\ z \in l\ \} \tag{3.16}$$

This matrix consists of two 25x3 blocks, one for each task. The compression along relevant dimensions $compr^A_{rel}$, $compr^B_{rel}$ and irrelevant dimensions $compr^A_{irrel}$, $compr^B_{irrel}$ as well as the context offset parameter $context_offset$ are multiplied with the respective blocks in this feature matrix:

$$X = \begin{bmatrix} X_A \\ X_B \end{bmatrix} = \begin{bmatrix} c_A & (1-compr^A_{rel})\, b_A & (1-compr^A_{irrel})l_A \\ context_{offset} * c_B & (1-compr^B_{irrel})b_B & (1-compr^B_{rel})l_B \end{bmatrix} \quad (3.17)$$

The rotation parameter determines the extent to which the representation of the first task is rotated into the frame of reference of the second task:

$$R_A(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & sin(\theta) \\ 0 & -sin(\theta) & cos(\theta) \end{bmatrix} \quad (3.18)$$

This rotation was applied to X_A in a subsequent step, so that the full model was given by:

$$X_{param} = \begin{bmatrix} [c_A & (1-compr^A_{rel})b_A & (1-compr_{irrel}A)l_A]R(\theta) \\ context_{offset} * c_B & (1-compr^B_{irrel})b_B & (1-compr^B_{rel})l_B \end{bmatrix} \quad (3.19)$$

We fit RDMs derived from this model to hidden layer RDMs using a constrained optimisation procedure with a least-squares cost function. As the procedure is sensitive to the choice of starting values, we averaged over 1000 independent runs with random starting values. We then performed group-level inference on the distribution of best-fitting parameter values, where the overall compression index was defined as the log of the ratio between compression along the relevant and irrelevant dimensions:

$$compression = log\left(\frac{compr_{rel}}{compr_{irrel}}\right) \quad (3.20)$$

Embedding dimensionality of hidden layer activity patterns : We used SVD to investigate the embedding dimensionality of the hidden layer activity patterns (Jazayeri & Ostojic, 2021). SVD was applied to the stimulus-by-unit matrix of hidden layer responses to all combinations of feature values and context. We visualised the cumulative variance explained based on the squared singular values (i.e., the eigenvalues of the response matrix) as Scree plot and performed the Elbow method to obtain a quali-

tative estimate of the embedding dimensionality. Next, we performed truncated SVD to assess the task-diagnosticity of the first k directions of variation in the response matrix. For this, we reconstructed the hidden layer response matrix, keeping only the first k singular values with k ranging from 1 to 27 (i.e. the number of input units). We then generated new outputs from the network by passing this lower-dimensional activity pattern on to the output unit. Lastly, we calculated the accuracy as the mismatch between these outputs and the ground truth, using the formula described in the section above. This allowed us to assess, separately for the rich and lazy regime, the extent to which removing components from the hidden layer responses reduced the network's performance. The hypothesis was that more components would be needed in the lazy compared to the rich regime to maintain equal task accuracy.

Hidden unit selectivity and axis alignment: To investigate task selectivity of hidden layer units, we capitalised on the property of ReLU nonlinearities that they map negative inputs to zero. We defined task-selectivity for the neural network as a non-zero response to stimuli in one context and zero response to all stimuli in the other context. Stimulus selectivity irrespective of context was defined as having a non-zero response in both contexts. We calculated these sensitivity indices at initialisation and after training to ensure that the initialisation scheme did not pre-partition the hidden layer in the absence of a training objective. Dead units were defined as returning zero for all stimuli (all combinations of feature values and context). From this, we calculated the proportion of units that were either dead, task- or stimulus-selective. To visualise response profiles, we averaged activity within these sub-populations, constructed a response matrix of these averages separately for each context (with rows corresponding to y location, columns to x-locations of stimuli and the value corresponding to the average activity of a sub-population) and plotted the group level average (mean across independent runs) as heatmaps. For this, we focussed on the two most extreme weight initialisations, 0.01 and 3, corresponding to learning in the rich and lazy regime, respectively. Lastly, to quantify the extent to which these response patterns were axis aligned (i.e., whether units responded to relevant but not irrelevant dimensions), we concatenated the two vectorised task response matrices, constructed RDMs based on

pairwise differences in magnitude and regressed them against two model RDMs, (1) the factorised and (2) linear models. In the factorised model, unit responses scaled with context-dependent relevant dimensions (i.e., with x-location in context A and y-location in context B). In the linear model, activity scaled jointly with both dimensions irrespective of context. We fitted the model at the level of individual runs. To assess which model RDM covaried stronger with the observed neural responses, we performed a Wilcoxon Signed Rank test on the difference between beta estimates for the factorised and linear model. To assess whether this difference was dependent on the initialisation scheme, we performed the same test on the difference of differences.

Context weight correlations: Our theory predicted that the network could learn the gating scheme via anti-correlated context weights. To test this empirically, we calculated the Pearson correlation between task A and task B weights from the input to the hidden layer at the level of single runs both at initialisation and after the last training epoch. We repeated this analysis on the sub-populations of task-specific and task-agnostic units, expecting weights into the former to be stronger anti-correlated. We visualised the distribution of single-run correlation coefficients together with a Kernel-Density-Estimate computed with the kdensity function from the Seaborn package.

Ablation study: We performed an ablation study to investigate how critical task-sensitive and stimulus-sensitive units were for multi-task performance. More specifically, for each collected run, we set either the sub-population of task-specific or task-agnostic units to zero, performed a forward pass through the ablated network and computed its loss and accuracy.

Chapter 4

Orthogonal task representations for context-dependent processing

Abstract

Humans can learn to perform multiple tasks that require to judge the same kind of stimuli according to different criteria. The neural geometry supporting this context-dependent continual task performance, however, is only poorly understood. Here we trained human participants on a context-dependent decision task and recorded fMRI data during a subsequent test phase. Representational Similarity Analysis of neural activity patterns revealed that the brain encodes information in a highly task-specific format, where task-relevant information is maintained and irrelevant information attenuated, forming a task representation with orthogonal coding axes. In a re-analysis of a freely available dataset with recordings from macaque FEF, we found evidence for a similar coding scheme. Together, these findings support a theory of neural coding according to which the prefrontal cortex attenuates task irrelevant information in a flexible, context-dependent manner, presumably to minimise interference between tasks.

4.1 Introduction

In the previous chapter we introduced a computational framework to study solutions to the context-dependent decision problem. We reported that neural networks could solve the task in two different ways, either by learning highly task-specific represen-

tations, where task-relevant information was mapped onto orthogonal axes, and irrelevant information filtered out, or with task-agnostic representations that recapitulated the structure of the input space. But how is this information represented in real brains?

One recently popular theory proposes that stimulus and context signals are projected into a high-dimensional neural code, permitting linear decoding of exhaustive combinations of task variables (Fusi et al., 2016). Indeed many neurons, especially in prefrontal and parietal cortex, exhibit nonlinear mixed selectivity, multiplexing information over several potentially relevant task variables (Raposo et al., 2014; Rigotti et al., 2013; Tang et al., 2019). This high-dimensional random mixed selectivity offers great behavioural flexibility because it maximises the potential for discrimination among diverse combinations of inputs, but also implies that neural codes should be relatively unstructured and task-agnostic. An alternative theory states that neural representations are mixed-selective but structured on a low-dimensional and task-specific manifold (Chaudhuri et al., 2019; Cueva et al., 2020; Ganguli et al., 2008; Gao & Ganguli, 2015; Sadtler et al., 2014), with correlated patterns of firing conferring robustness on the population code (Zohary et al., 1994). Representations may adapt so that irrelevant task information is wholly or partially filtered out in ways that minimise interference between tasks (Cohen et al., 1990; Miller & Cohen, 2001), consistent with accounts emphasising that neural codes are sculpted by task demands (Duncan, 2001) or through attention to scenes and objects (Cukur et al., 2013). The question of whether neural codes are task-agnostic or task-specific speaks to core problems in neural theory with widespread implications for understanding the coding properties of neurons and neural populations (Saxena & Cunningham, 2019; Yuste, 2015).

Comparably little is known about task representations in humans, as most of the previous work has focussed on non-human primates (NHPs) (Badre et al., 2021). To address this gap in the literature, we studied the geometry and dimensionality of task representations recorded with fMRI in human participants trained on a paradigm involving the classification of high-dimensional images of fractal trees. We found evidence for structured, low-dimensional and task-specific representations in the fronto-parietal network, resembling those we had previously observed in neural networks

trained in the rich learning regime. Capitalising on a freely available dataset with recordings from macaque FEF (Mante et al., 2013) enabled us to test more nuanced predictions from the gating theory proposed in the previous chapter. Like in our human participants, we found evidence for structured, low-dimensional representations in the population activity. Looking at the response profiles of individual neurons revealed that the majority of units coded for the task-relevant (but not irrelevant) feature dimension in a context-dependent manner, as predicted by the gating theory. Some key differences between the experimental designs, however, might permit alternative interpretations of the NHP data, which we extensively review in the discussion.

4.2 Results

We focus on a canonical paradigm involving context-dependent classification of D-dimensional stimuli $x(i,j) \in \mathbb{R}^D$ which vary along two underlying dimensions i and j, for which correct decisions depend on i in task A and j in task B. Healthy human participants ($n = 32$) categorised naturalistic (tree) stimuli, with the correct class given by branch density in one context and leaf density the other (**Fig. 4.1A,B**). We varied these two dimensions parametrically to generate an n-by-n grid of unique stimuli in which the density of branches and leaves were independent by design. The dimensions were a priori unknown to participants (Flesch et al., 2018).

4.2.1 Behavioural results indicate that participants learned accurate estimates of the true category boundaries

Accuracy increased with training, jumping from $64 \pm 2\%$ to $88 \pm 2\%$ between an initial baseline and a final test conducted in the fMRI scanner ($T(29) = 11.1, p < 0.001$, **Fig. 4.1C**). Using a psychophysical model to decompose errors into distinct sources (see methods), we found that this improved performance was due neither to a steepening of the psychometric curve (*slope:* $p = 0.120$), nor a reduction in decision bias (*offset:* $p = 0.319$) although the scan session was characterised by fewer generic lapses (*lapse:* $Z = -3.5, p < 0.001$, **Fig. 4.1E**). Instead, the fitted estimation error for the category boundary fell from 27 degrees to 7 degrees (*angular bias:* $Z = -4.1, p < 0.001$, **Fig. 4.1E**). In a previous study (Flesch et al., 2018) we quantified behavioural response

patterns in this trees task by fitting a model that made choices according to the two orthogonal ground truth boundaries. This factorised model fit better than a linear model that learned a single boundary for both tasks, a finding we replicate here (**Fig. 4.1F**; *scan phase: factorised > linear* $T(29) = 17.61, p < 0.0001$, *phase x model interaction:* $T(29) = -10.84, p < 0.0001$). In other words, despite having no prior knowledge of the tasks or stimulus space, participants learned over the course of training to apply the orthogonal category boundaries appropriately in each context (**Fig. 4.1D**).

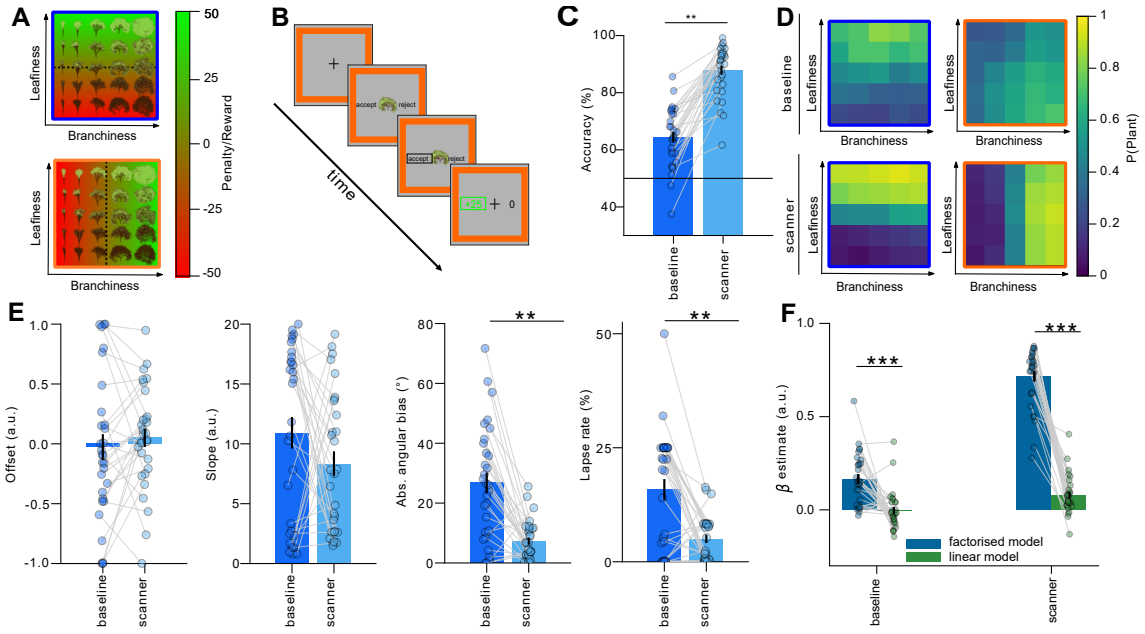

Figure 4.1: Task design and behavioural findings. (**A**) Illustration of the 2D stimulus space. Each image shows the category boundary (dashed line) and reward/penalty (red-green colour) for choosing to plant in a specific context (signalled by blue frame/orange frame). (**B**) Example trial sequence. Participants were asked to "accept" (plant) or reject a tree by pressing one of two buttons. Frame colour signalled context. Participants received rewards and penalties for planting trees. (**C**) Mean accuracy improved from baseline to scan. Each dot is a participant. (**D**) Choice matrices show the mean probability of choosing "plant" for each tree (defined by a level of leaf/branch density) in each context, for both the baseline (top) and scanner (bottom) sessions. (**E**) Parameters of the psychophysical model between baseline and scan: offset, slope, angular bias, and lapse rate. Each dot is a participant. ** denotes $p < 0.01$. (F) Fits of linear and factorised model at baseline and scan. Each dot is a participant. Error bars indicate SEM.

4.2.2 Univariate effects in BOLD signal replicate standard effects from the literature

Before we investigated the neural geometry of task representations, we carried out a series of univariate analyses of the BOLD signal to test whether we could replicate standard findings from the literature. A well-established result in task-switching paradigms is that performance tends to be lower on task-switch compared to task-stay trials, indicated by differences in task accuracy and reaction time (Monsell, 2003). Indeed, participants were slightly worse on switch than stay trials at test, both during the baseline and later scanning session (*Accuracy Baseline, Switch < Stay:* $T(29) = 2.057, p = 0.048, d = 0.266$*; Accuracy Scan, Switch < Stay:* $T(29) = 2.715$*,* $p = 0.011$*,* $d = 0.211$*; Interaction Phase x Switch cost:* $T(29) = -0.668$*,* $p = 0.509$*,* $d = -0.251$*)*, **Fig. 4.2A**). At a neural level, this switch cost is associated with heighted BOLD signals in prefrontal areas on switch compared to stay trials (Yeung et al., 2006). A whole-brain univariate contrast of switch versus stay trials revealed clusters in task-positive regions where activity was higher on switch than on stay trials. More specifically, we found significant clusters in Parietal Cortex (*BA7 :* $T(30) = 5.65$*,* $p < 0.001$ *(FWE corrected), cluster extent* $(kE) = 570$*, MNI coordinates* = $[-6, -74, 52]$), Supplementary Motor Area (*SMA:* $T(30) = 5.03$*,* $p < 0.05$*,* $kE = 66$*,* $[-6, 18, 46]$) and left Medial Frontal Gyrus (*MFG:* $T(30) = 6.55$*,* $p < 0.01$*,* $kE = 124$*,* $[-44, 21, 28]$) (**Fig. 4.2B**). Another common finding is that the BOLD signal scales with the amount of decision certainty (Tosoni et al., 2008). In our paradigm, this was modulated by the distance of a stimulus to the category boundary. Fitting logistic functions to the choice patterns along both dimensions revealed that, compared to the baseline, participants became much more sensitive to the task-relevant dimension after they had engaged in the blocked training phase (*Slope Relevant, Baseline:* $Z = 4.72, p =< 0.001, d = 0.873$*; Scan > Baseline:* $Z = 4.762, p =< 0.001$*; Scan:* $Z = 2.705, p = 0.007, d = 0.494$). Participants were, however, much more sensitive to the relevant than irrelevant dimension at test (*Scan, Relevant > Irrelevant:* $Z = 4.782, p < 0.001, d = 0.873$), and this sensitivity was higher compared to baseline (*Dimension x Phase Interaction:* $Z = 4.741, p < 0.001, d = 0.866$) (**Fig. 4.2C**). A GLM with parametric regressors

for the absolute distance to category boundary (methods) revealed significant relationships between activity and distance to bound along the relevant, but not irrelevant feature dimensions. For the relevant dimension, we found significant clusters in Angular Gyrus (*left:* $T(30) = 6.79,\ p < 0.001,\ kE = 364,\ [60,-49,28]$), the right Orbitofrontal Corex ($T(30) = 5.46,\ p < 0.01,\ kE = 73,\ [8,42,-14]$), and to a lesser extent also in bilateral EVC (*left:* $T(30) = 5.15,\ p < 0.01,\ kE = 70,\ [-13,-98,14]$; *right:* $T(30) = 6.55,\ p < 0.01,\ kE = 61, [18,-94,21]$) as well as the Posterior Cingulate cortex ($T(30) = 5.05,\ p < 0.001,\ kE = 192,\ [4,-49,35]$) (**Fig. 4.2D**). Finally, we looked at univariate markers of choice value, which have previously been reported in prefrontal regions (Boorman et al., 2011). A GLM with regressors for the choice and value of the stimuli revealed significant relationships between the interaction of choice and value and BOLD. Consistent with previous reports, we found clusters in ACC ($T(30) = -9.52,\ p < 0.001$ uncorr , $kE = 803,\ [8,28,31]$), VMPFC ($T(30) = 4.33$, $p < 0.001$ uncorr, $kE = 23$, $[4.5, 38.5, -21]$) and the striatum ($T(30) = -8.88$, $p < 0.001$, $kE = 302$, $[-2.5, -14, -3.5]$) (**Fig. 4.2E**). Together, these findings demonstrate that classic effects from the task switching and decision-making literature replicate in a paradigm where participants learned context-dependent decisions with high-dimensional stimuli solely from trial-wise feedback.

4.2.3 Human fMRI reveals task specific representations consistent with those predicted by the rich training regime

How are task representations structured in biological brains? The simulations presented in the previous chapter furnished predictions about the neural geometry we should expect to see in BOLD data acquired during the final phase of our experiment. To investigate this neural geometry, we once again turned to a more powerful multivariate analysis of the activity patterns (RSA). We used model RDMs encoding grid, orthogonal, parallel, and various control patterns to predict brain activity using a spherical searchlight across the whole brain (**Appendix B, Fig. B.2**). Crucially, we observed strong correlations with the orthogonal model in three major foci: the dorsolateral prefrontal cortex (*dlPFC:* $T(30) = 9.79,\ p < 0.001$ *corrected, peak* $[46,14,24]$), the mid-cingulate cortex (*MCC:* $T(30) = 9.51,\ p < 0.001$ *corrected,*

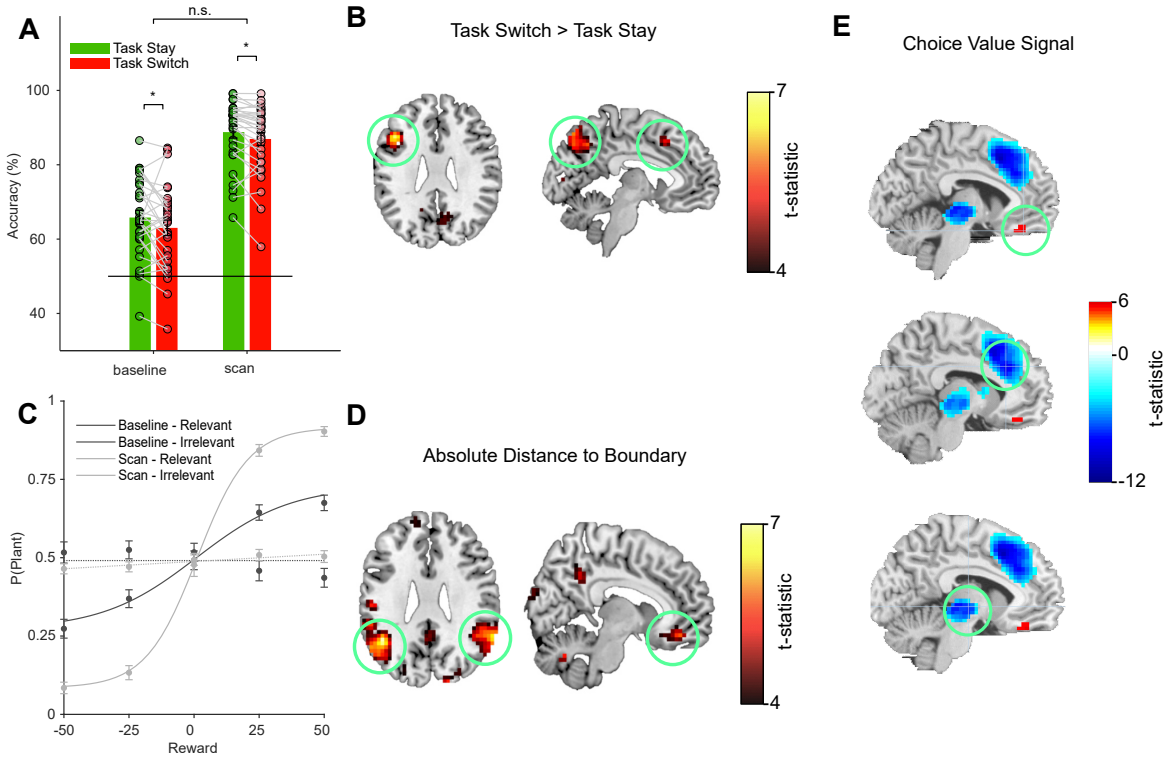

Figure 4.2: Replication of standard behavioural and univariate findings. (**A**) Behavioural switch cost. Accuracy was higher on stay compared to switch trials. While participants performed significantly better in the scanning compared to the baseline session, the difference in switch cost was not significant. (**B**) Neural correlates of switch cost, showing regions where BOLD activity was higher on switch compared to stay trials. (**C**) Psychometric curves for relevant and irrelevant dimensions. Participants got more sensitive to the relevant dimension with training. Overall, there was no evidence for intrusions from the irrelevant dimensions. (**D**) Neural correlates of absolute distance to category boundary, showing regions where the BOLD signal scaled significantly with the absolute distance to boundary. (**E**) Neural correlates of a choice value signal, encoding the interaction between choice (accept/reject) and the value of the chosen stimulus. All error bars denote SEM.

peak $[8,21,49]$) and the posterior parietal cortex (*PPC:* $T(30) = 8.87$, $p < 0.001$ *corrected, peak* $[39,-45,45]$), **Fig. 4.3A**). A similar effect was observed in a left prefrontal region for which the univariate analysis had revealed that it was sensitive to task switches, but the fit of the orthogonal model did not differ between switch and stay trials (**Appendix B, Fig. B.2**). In early visual regions, neural data RDMs were best predicted by a model in which dissimilarities depended mainly on branch density ($T(30) = 6.98$, $p < 0.001$ corrected, peak $[22,-84,-3]$) but no other models explained a significant amount of variance in the neural RDMs (**Fig. 4.3A**). Repeat-

ing this RSA with independently defined ROIs confirmed that the branchiness model fit best in EVC (*Bonferroni-corrected* $\alpha = 0.0071$*; Grid:* $T(30) = 3.46, p = 0.0008$*; Rotated Grid:* $T(30) = 0.93, p = 0.1809$*; Orthogonal:* $T(30) = 1.76, p = 0.0442$*; Parallel:* $T(30) = -0.75, p = 0.7692$*; Branchiness:* $T(30) = 4.74, p < 0.0001$*; Leafiness:* $T(30) = -3.43, p = 0.9991$*; Diagonal:* $T(30) = -1.20, p = 0.8805$, **Fig. 4.5A**), whereas the orthogonal model fit best in DLPFC (*Grid:* $T(30) = 0.60, p = 0.2758$*; Rotated Grid:* $T(30) = -0.19, p = 0.5746$*; Orthogonal:* $T(30) = 8.18, p < 0.0001$*; Parallel:* $T(30) = -1.31, p = 0.8999$*; Branchiness:* $T(30) = 0.19, p = 0.4259$*; Leafiness:* $T(30) = -1.59, p = 0.9388$*; Diagonal:* $T(30) = -1.28, p = 0.8949$), PPC (*Grid:* $T(30) = 1.31, p = 0.1007$; *Rotated Grid:* $T(30) = -0.56, p = 0.7086$*; Orthogonal:* $T(30) = 7.77, p < 0.0001$*; Parallel:* $T(30) = -1.41, p = 0.9149$*; Branchiness:* $T(30) = -0.35, p = 0.6354$*; Leafiness:* $T(30) = -2.19, p = 0.9816$*; Diagonal:* $T(30) = -2.73, p = 0.9947$) and MCC (*Grid:* $T(30) = 1.08, p = 0.1448$*; Rotated Grid:* $T(30) = 0.82, p = 0.2093$*; Orthogonal:* $T(30) = 7.17, p < 0.0001$*; Parallel:* $T(30) = -1.88, p = 0.9648$*; Branchiness:* $T(30) = -0.50, p = 0.6908$*; Leafiness:* $T(30) = -1.26, p = 0.8914$*; Diagonal:* $T(30) = -2.18, p = 0.9815$, **Fig. 4.5A**). To verify that EVC encoded both dimensions irrespective of the task, whereas frontoparietal regions employ partially compressed and orthogonal representations, we fit a Support Vector Machine (SVM) with linear kernel and binary outputs to the relevant feature dimensions (high vs low branchiness/leafiness) in each region and assessed the cross-validated decoding performance along relevant and irrelevant dimension in the same and other task. In all regions, decoding accuracy along the relevant dimension of the task the decoder had been trained on was significantly above chance (*Bonferroni corrected* $\alpha = 0.0125$*; EVC:* $T(30) = 8.99, p < 0.0001$*, DLPFC:* $T(30) = 4.41, p = 0.0001$*, MCC:* $T(30) = 5.54, p < 0.0001$*, PPC:* $T(30) = 4.13, p = 0.0003$, **Appendix B, Fig. B.4**). The same dimension in the other task could be reliably decoded in EVC but not in the other regions, again suggesting that those regions attenuated irrelevant dimensions relative to the relevant ones (*Bonferroni corrected* $\alpha = 0.0125$*; EVC:* $T(30) = 8.44, p < 0.0001$*, DLPFC:* $T(30) = 2.32, p = 0.0275$*, MCC:* $T(30) = 1.22, p = 0.232$*, PPC:* $T(30) = 1.85, p = 0.0736$, **Appendix B,**

Fig. B.4). To statistically compare representations in these regions, we conducted a Bayesian model comparison (RFX BMS) of linear regressions with and without the branchiness/orthogonal model RDM, fit separately to EVC, DLPFC, PPC and MCC. Protected exceedance probabilities (pep) that quantify how likely it was that the same model explained patterns in EVC and DLPFC/PPC/MCC were extremely low (*EVC & DLPFC:* $pep = 0.000329$*, EVC & MCC:* $pep = 0.0029$*, EVC & PPC:* $pep = 0.000191$). RFX BMS within each region confirmed again that the branchiness model explained most of the patterns in EVC, while the orthogonal model yielded the best fit in DLPFC/MCC/PPC (**Fig. 4.5; Appendix B, Table B.8**). To summarise, in fronto-parietal areas, neural codes were largely structured as predicted by rich learning, with representations in each context projected onto orthogonal neural axes that are elongated along the relevant feature dimension and compressed along the irrelevant feature dimension. In contrast, representations in early visual areas were largely unaffected by context.

4.2.4 Fronto-parietal representations in the human brain are task-specific and low-dimensional

Next, we fit the parametric RSA model to the neural data within each independently defined ROI to quantify the extent to which irrelevant information was attenuated in each context. This confirmed that in DLPFC/PPC/MCC, the neural code was compressed along irrelevant relative to relevant dimensions and remained in the naïve (input) space rather than being rotated into the frame of reference of the response ("accept" vs "reject" irrespective of context) (*EVC Compression Leafiness Task:* $z = 2.25, p = 0.0242$*, Compression Branchiness Task:* $z = 4.10, p < 0.0001$*, Offset:* $z = 4.86, p < 0.0001$*, Rotation:* $z = 1.49, p = 0.1364$*, DLPFC Compression Leafiness Task:* $z = 4.53, p < 0.0001$*, Compression Branchiness Task:* $z = 4.86, p < 0.0001$*, Offset:* $z = 4.86, p < 0.0001$*, Rotation:* $z = 1.14, p = 0.2557$*, MCC Compression Leafiness Task:* $z = 4.84, p < 0.0001$*, Compression Branchiness Task:* $z = 4.80, p < 0.0001$*, Offset:* $z = 4.86, p < 0.0001$*, Rotation:* $z = 0.27, p = 0.7838$*, PPC Compression Leafiness Task:* $z = 4.80, p < 0.0001$*, Compression Branchiness Task:* $z = 4.84, p < 0.0001$*, Offset:* $z = 4.86, p < 0.0001$*, Rotation:* $z = 0.53, p = 0.5967$, **Fig.4.3B**). When we

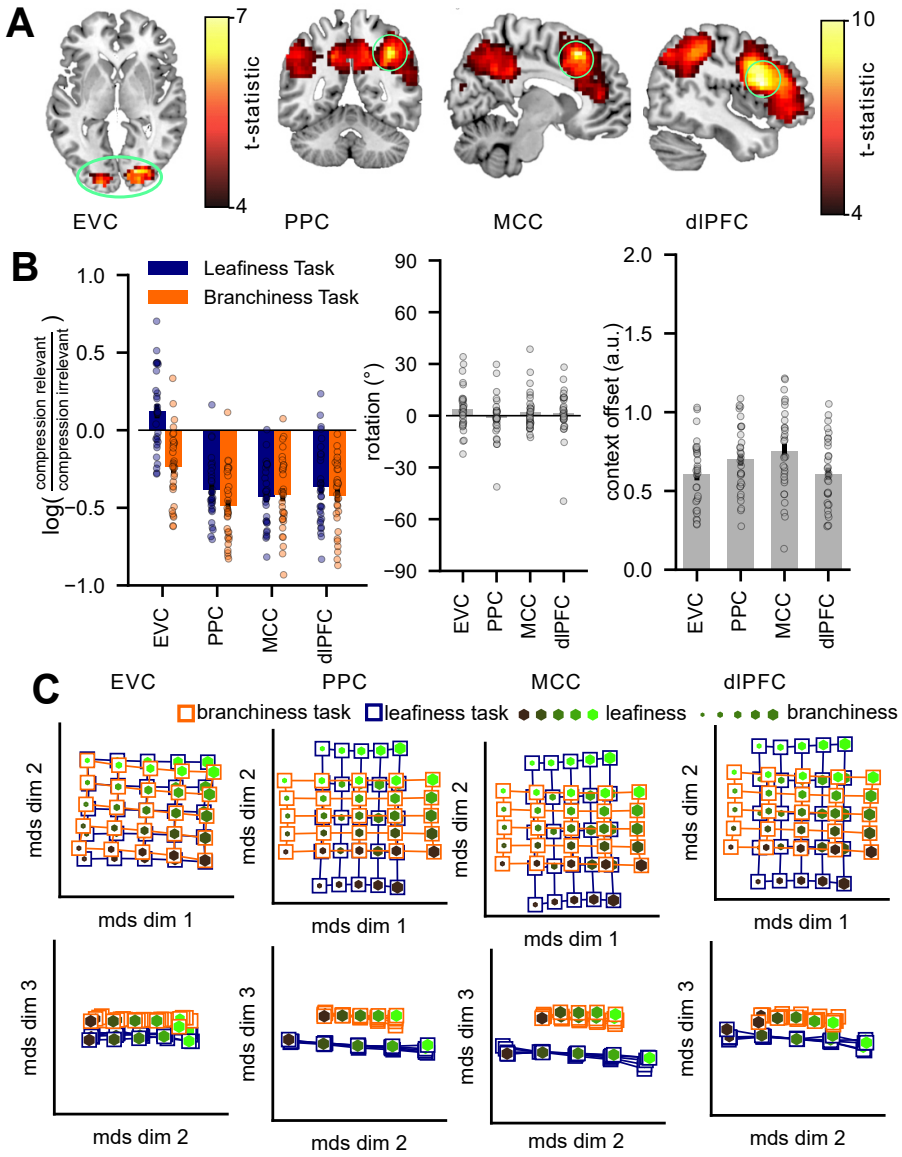

Figure 4.3: Evidence for task-specific representations in human fMRI data. (**A**) Results from searchlight RSA. Left panel: voxel regions where neural similarity patterns matched the branchiness RDM. Right panel: voxels where neural similarity patterns matched the orthogonal RDM. All data are corrected for multiple comparisons. (**B**) Data from parametric RDM fits. (**C**) Low-dimensional projections of fMRI data from within ROIs taken from visual, parietal and frontal regions, reconstructed from coefficients of regression model described in (B).

used MDS to visualise the best-fitting model RDMs for each region in three dimensions, the task-specific encoding of relevant dimensions along orthogonal manifolds in dorsal stream regions of interest can be clearly seen (**Fig. 4.3C**). Finally, in neural networks rich learning is characterised by a low-dimensional neural code. Interest-

ingly, PCA on the neural data suggested that patterns in fronto-parietal regions were higher dimensional than in EVC (number of PCs needed to explain 95% of variance (*EVC/DLPFC/PPC/MCC: 9/15/19/19*, **Appendix B, Fig. B.5A**). However, by systematically removing components from the data using PCA on the BOLD patterns within each candidate ROI and repeating the RSA on this reduced space, we revealed that reliable correlation with the orthogonal manifolds RDM required just two components in each region of interest and that there was no measurable benefit in maintaining more than 6 PCs in total (**Appendix B, Fig. B.5B**). Similar patterns were seen for the grid model in EVC (**Appendix B, Fig. B.5B**). In other words, the neural representations seem to be embedded in a low-dimensional subspace focused on task-relevant stimuli, as predicted by rich learning.

4.2.5 Neural task factorisation predicts behavioural axis alignment

Next, we attempted to link these neural patterns to behaviour. In theory, if participants had learned to filter out irrelevant dimensions, one would expect that their category judgements showed fewer signs of intrusions from those dimensions. The factorised model that was fit to human choices quantified the extent to which these choices were aligned with the ground truth category boundaries. This yielded an "axis alignment" score for each participant, which was correlated with the orthogonality of neural task representations across the cohort in PPC (*Kendall's* $\tau_a = 0.27, p = 0.038$), MCC (*Kendall's* $\tau_a = 0.36, p = 0.005$) and dlPFC (*Kendall's* $\tau_a = 0.38, p = 0.003$; **Fig. 4.4**). In other words, the category judgements of participants with more factorised neural representations respected more orthogonal category boundaries, suggesting a link between the extent to which task information is embedded in orthogonal manifolds and the ability to avoid mutual interference between tasks.

4.2.6 Representations in NHP single-unit data are consistent with predictions from rich learning

BOLD data offers at best an indirect window on neural coding, so we additionally capitalised on a freely available dataset describing single neuron activity in frontal

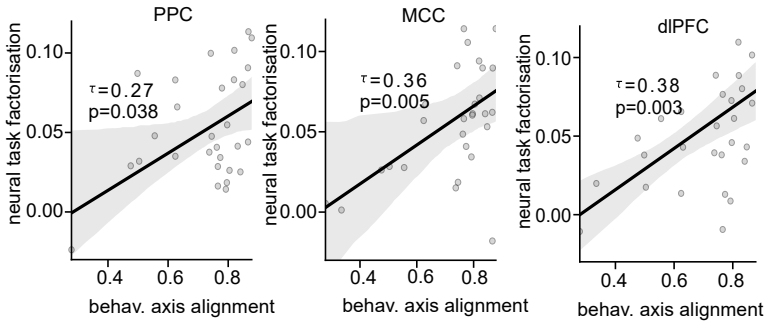

Figure 4.4: Brain behaviour correlations Correlation between neural task factorisation (fits of orthogonal model to neural data) and behavioural axis alignment (fits of factorised model to choice matrices). Each dot is a participant.

eye fields (FEF) whilst macaques performed an equivalent context-dependent decision task on stimuli with varying colour and motion coherence (Aoi et al., 2019; Mante et al., 2013). We focus on the results from monkey A, because our analyses (and those reported previously) indicate that FEF neurons recorded from monkey F were only very weakly sensitive to colour even when it was decision-relevant (Aoi et al., 2019) (**Appendix B, Fig. B.6C-D**). First, we built a pseudo-population from all the recorded neurons and visualised its neural geometry in 2 dimensions with MDS. This revealed two orthogonal manifolds, each coding for one of the two task-relevant axes, like the one observed in BOLD data and predicted by neural networks trained in the rich regime (**Fig. 4.6A**). Indeed, when we fit the candidate RDMs used above to the dataset, the orthogonal RDM fit best for monkey A (*grid model:* $p = 0.027$*, orthogonal model:* $p < 0.0001$*, only motion model:* $p = 0.006$, **Fig. 4.6B**); an RDM coding for motion alone fit best for monkey F (*grid model:* $p = 0.004$*, orthogonal model:* $p < 0.0001$*, only motion model:* $p < 0.0001$, **Appendix B, Fig. B.6C**). Training a linear SVM on the patterns recorded in monkey A confirmed that the relevant, but not the task-irrelevant dimension could be reliably decoded (*same task, rel. dim.:* $p < 0.0001$*; other task, irrel. dim.:* $p = 0.064$; **Appendix B, Fig. B.6A**). We also tested dimensionality of these neural geometries using a similar PCA-based approach as above; the ability to decode orthogonal manifolds dropped sharply when fewer than 3 components were retained, suggesting that directions of highest variance were aligned with the task-relevant dimensions of context, color, and motion (**Appendix B, Fig. B.6B**). This

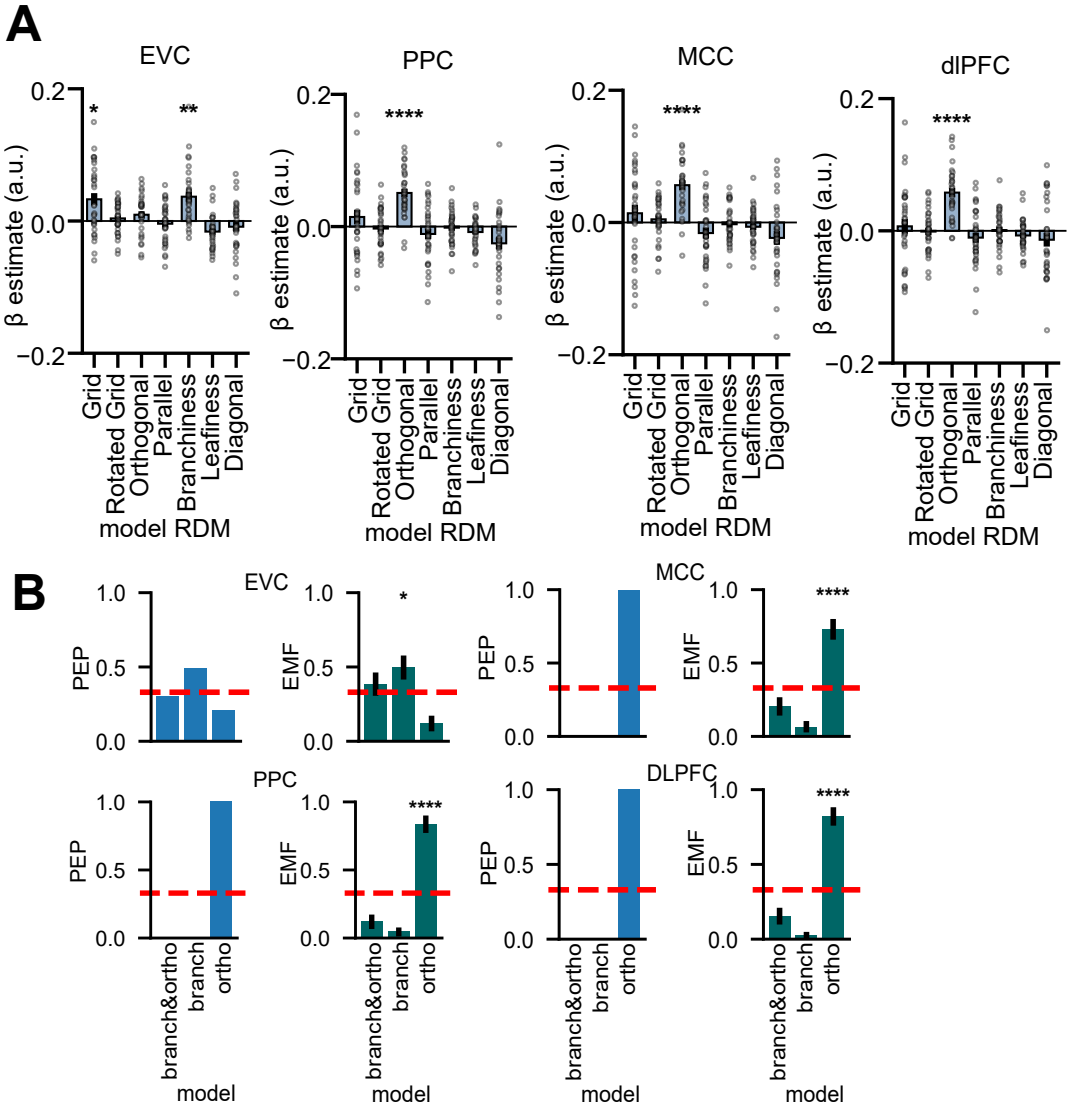

Figure 4.5: Further RSA results for human fMRI data. (**A**) Results from fitting the seven model RDMs to four independently defined ROIs, showing task-agnostic encoding in EVC and task-specific encoding in fronto-parietal areas. (**B**) Protected exceedance probabilities (left) and estimated model frequencies (right) of a Bayesian model comparison between rich and lazy RDMs for the four candidate regions. Protected exceedance probabilities of comparisons between regions (not shown) implied that it was very unlikely that the same model explained patterns in EVC and DLPFC/PPC/MCC (EVC & DLPFC: pep=3.29e-4, EVC & MCC: pep=0.0029, EVC & PPC: pep=1.91e-4). RFX BMS within each region confirmed again that the branchiness model explained most of the patterns in EVC, in contrast to the orthogonal model in DLPFC/PPC/MCC. Full statistical results are reported in Appendix B, Table B.8.

analysis suggests that the orthogonal manifolds identified with the RSA lie embedded in a very low-dimensional manifold and indicates that the effect observed in human

BOLD may generalise across species and recording methods.

4.2.7 Single-unit task selectivity in NHP data supports gating theory

The high spatial resolution of the non-human primate (NHP) recordings allowed us to test more fine-grained predictions from the gating theory presented in the previous chapter. Under rich learning, most units should code for the relevant feature in a context-dependent manner. Indeed, in the NHP data, we found that the majority (65%) of significantly responsive units were also selective to either colour in the colour task or motion in the motion task, although there was strong bias towards the motion task (**Fig. 4.6C**). Looking at the activity profiles of these units in detail revealed that responses of task-specific units were aligned to the two choice axes (**Fig. 4.6D-E**, *factorised model > linear model:* $z = 4.643, p < 0.0001, d = 0.558$). By contrast, those units without coding preference for either task coded for a residual policy which collapses across both contexts ("task agnostic"), resembling the linear model described above (**Fig. 4.6E**, *linear model > factorised model:* $z = 4.033, p < 0.0001, d = 0.749$). Together, these findings support the theory that task-relevant information is gated into orthogonal subspaces in a way that minimises mutual interference.

4.3 Discussion

In this chapter, we investigated the neural geometry of task representations in human participants trained on a context-dependent decision-making problem. We found strong evidence for highly task-specific representations in fronto-parietal networks, where irrelevant feature dimensions were attenuated, so that information was encoded on orthogonal manifolds, with different axes coding for different tasks. Complementary analyses of a freely available dataset with recordings from macaques who had completed a similar task revealed remarkable parallels between the representations observed in both species and allowed us to test more fine-grained predictions of the gating theory outlined in the previous chapter. Together, these results support a theory in which prefrontal cortex learns to gate out task-irrelevant information and project

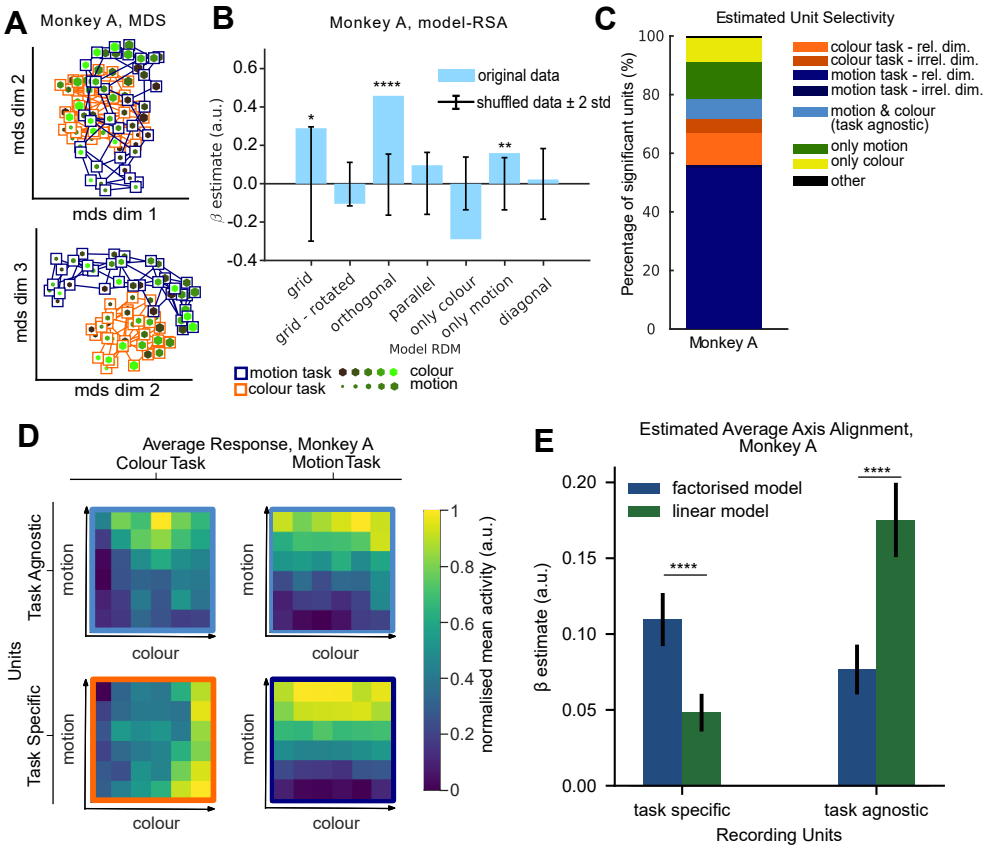

Figure 4.6: Evidence for orthogonal representations in recordings from Macaque FEF. (**A**) MDS projection of data from monkey A. Stimulus features are now colour and motion; data from (Mante et al., 2013). (**B**) Fits of orthogonal, grid and control RDMs to data from monkey A. Error bars indicate +- 2std of RSA on shuffled data (1000 permutations). (**C**) Distribution of selectivity of single units in monkey A, showing that the majority of units was task-selective. (**D**) Selectivity of recording units from monkey A for each relevant and irrelevant stimulus feature in each context. Note that task-specific units (lower panels) are mostly sensitive to relevant vs. irrelevant dimension whereas task-agnostic units show tendency towards coding for an interaction of both features (diagonal). (**E**) Quantification of results in (D) using fits of linear vs. factorised model. The factorised model fits best to task-specific units, and the linear model to task-agnostic units.

relevant information on orthogonal coding axes, presumably to minimise between-task interference (Cohen et al., 1990; Musslick et al., 2017).

This is quite striking, because conflicting reports have suggested that task-irrelevant information is retained or discarded during context-dependent decision-making (Mante et al., 2013; Takagi et al., 2020). Previous analyses of single cell data from macaque prefrontal cortex have emphasised that neural selectivity is mixed, and

representations are high dimensional, in seeming contradiction to the findings reported here (Fusi et al., 2016; Rigotti et al., 2013). One possibility is that over prolonged training, the dimensionality and geometry of neural representations is tailored to the transfer demands of the paradigm (Musslick et al., 2017). Structured, low-dimensional representations may be favoured in settings where information can be shared across tasks or stimuli, such as our trees task, where all stimuli were unique, but sampled from the same underlying generative process, hence permitting generalisation of latent features across tasks. By contrast, high-dimensional neural codes may emerge by preference in tasks with minimal need for generalisation, such as recall and recognition of a small set of unrelated images (Barak et al., 2013; Rigotti et al., 2013). Another possibility is that training duration and/or task instructions influence representational geometry. Our participants had to infer task-relevant dimension from trial-wise feedback alone and were given extensive training prior to the scanning session. Future studies could explore whether representations consistent with lazy learning would be observed when participants were made aware of the task-relevant dimensions a priori and/or received less training. This concern speaks to a broader question of how the human brain uses context to partition experiences, and how the specific training regime we subjected our participants to might have shaped neural representations. In previous work, we demonstrated that humans learn better under blocked curricula, where one task is learned at a time, compared to interleaved curricula where both tasks are randomly interspersed (Flesch et al., 2018). For the fMRI study, we gave our participants a blocked curriculum. Future work would be required to investigate how the temporal statistics of training samples shape neural geometry.

At first glance, some of our findings might appear to diverge from previous analyses of the same data, in that we emphasise that irrelevant information is at least partly compressed in FEF (Aoi et al., 2019; Mante et al., 2013). However, our analysis of the NHP data focussed on a relatively late epoch (300-600ms post-stimulus). In fact, when we repeated the model-based RSA separately for early, middle and late time windows following stimulus onset, we found that representations were more grid-like early on (encoding of both feature dimensions) but became highly task-specific in the second

half of the trial (**Appendix B, Fig. B.7C**). Crucially, we can explain this temporal evolution of task representations with an extension of our gating theory that incorporates recurrence into the neural network model (**Appendix B, Fig. B.7A**). Under this account, feature-selective units keep integrating motion/colour information throughout the stimulus presentation period, but the irrelevant dimension is integrated at a slower rate, giving rise to a gradual progression from grid-like to orthogonal representations. In the following delay-period, the context cue continues to act as inhibitory bias on the unit encoding irrelevant features, gradually surpressing its activity just enough so that by the time of a response, only task-relevant information is preserved, leaving a fully orthogonal and task-specific representation (**Appendix B, Fig. B.7B**). When we visualised the geometries separately for early, middle and late windows within the stimulus interval, we observed a similar temporal evolution from grid-like to more orthogonal representations in both the RNN and monkey recordings (**Appendix B, Fig. B.7D**). Notably, similar evolutions of representational geometries over the time course of a trial have recently been reported in the context of working memory tasks (Panichello & Buschman, 2021).

By applying the same analysis to the veridical stimulus space in different recording modalities, we found converging evidence for orthogonal task-specific representations in neural networks, fMRI recordings from humans and single-unit recordings from macaque FEF. However, it should be noted that the two latter studies differed in their way in which responses were counter-balanced across trials, which points towards an alternative interpretation. To prevent confounds with motor responses, we fully orthogonalized responses for both tasks of the fMRI study, by randomising the assignment of accept/reject to the two response buttons. In contrast, in Mante et al., (2013), this randomisation was only applied to the context in which colour was relevant. As macaques were instructed to respond with saccades either to the left or right side of the screen, and the random dots moved either to the left or right, this could have introduced a partial confound between the relevant information in one context and the saccadic response. Consistent with this view, we observed that more units were selective to motion in the motion context, compared to colour in the colour context . This

suggest an alternative interpretation of the data that would emphasize an encoding of the choice, indicated by leftward versus rightward saccades, rather than the veridical stimulus information (Mante et al., 2013). Consistent with this view are early reports suggesting that FEF's primary role lies in the preparation of saccadic responses (Robinson & Fuchs, 1969). Support for this hypothesis comes from our simulations with deeper networks, where we observed a progressive transformation from grid-like over orthogonal to parallel representations, as the signal propagated from the input to the output layer, suggestive of a gradual transformation of information from the frame of reference of the inputs to the frame of reference of the response. However, we explicitly tested for an encoding of the response (accept/reject) with our parallel model, and failed to find evidence for this in the macaque data at any timepoint of the trial.

One possibility is that this failure to detect such a signal is a by-product of the uneven counterbalancing of responses in this dataset. Another possibility is that the available recordings from FEF include information in both reference frames. Indeed, more recent evidence has contested the view that FEF is solely involved in saccade preparation. According to the premotor theory of attention, both covert and overt attention rely on the same neural substrates (Rizzolatti et al., 1987). Indeed, neuroscientific evidence suggests that FEF could be involved in both processes (Corbetta et al., 1998), and more fine-grained investigations have revealed sub-divisions into populations of neurons coding for a saliency map of task-relevant information, and those coding for saccades to relevant targets (Sato & Schall, 2003). Together, these results may suggest that top-down modulatory processes and saccade preparation could be functionally intertwined in FEF (Schafer & Moore, 2007), a perspective which could possibly accommodate both interpretations of the macaque data. Future work could apply our RSA-based analyses framework to recordings from different areas in the macaque brain, such as the dataset reported in (Brincat et al., 2018; Siegel et al., 2015), and test whether the functional distance of prefrontal recording sites to motor areas determines whether parallel or orthogonal geometries are observed.

Taken together, our findings suggest striking similarities between representations of task rules in biological and artificial neural networks. The results indicate that for

context-dependent decision tasks learned sequentially via trial and error, the human brain appears to utilise a coding scheme that minimises representational overlap between these tasks, like the one adopted by a neural network trained in the rich regime on interleaved data.

4.4 Methods

4.4.1 Participants

Human participants: A total of 32 participants (mean age 24.44y, 31 right-handed, 21 female) with no history of neurological or psychiatric disorders were recruited from a participant pool at the University of Granada. One participant was excluded from the analysis due to equipment failure during the scanning session, leaving 31 participants for the fMRI analysis. For another participant training data was not recorded due to disruption of their internet connection, leaving 30 participants for all behavioural analyses. All participants gave written informed consent prior to taking part in the study. The experiment received approval from the ethics board of the University of Granada. Participants were compensated for their time with 38€. The experiment consisted of several sessions completed on three successive days (**Appendix B, Fig. B.1A**). All participants completed a pre-screening study on day 1 that assessed their eligibility for the main experiment. The main experiment consisted of a browser-based training session on day 2, and a refresher and scanning session on day 3, which took place at the fMRI institute of the University of Granada.

Nonhuman primate data: NHP results were based on a reanalysis of data recorded from monkey frontal eye fields (FEF) during performance of comparable context-based decision-making tasks. These data have already been intensively scrutinised in past work (Aoi et al., 2019; Mante et al., 2013). In the experiment, two monkeys were asked to discriminate between distinct levels of motion direction and colour of random dot stimuli, with only one dimension being relevant in each context, just as in our experiments. Stimuli spanned a similar 2D grid (motion directions varying from left to right, colour gradient from green to red) as our trees and Gaussian blobs. Further details are available in ref (Mante et al., 2013).

4.4.2 Task Design

Stimuli: Participants performed a virtual gardening task for which they had to discover rules that determined growth success of tree stimuli in two different gardens. Trees were generated by in house-code and were built to vary parametrically in five discrete steps along two different dimensions, the density of leaves ("leafiness") and the density of branches ("branchiness"), yielding 25 unique class. We generated multiple stimuli per level of leafiness and branchiness and sampled these exemplars randomly without replacement for training and test sessions at the level of individual participants so that no physical stimulus was presented twice during the experiment.

Pre-Screening Session (Day1): We previously showed that learning is mediated by an a priori tendency to factorise tree space into dimensions of leafiness and branchiness (Flesch et al., 2018). To measure this prior in our participants we first used an online task in which participants moved tree exemplars within a circular open arena via drag and drop on the screen, attempting to arrange them so that distance between trees was proportional to their perceived dissimilarity (**Appendix B, Fig. B.1B**). Participants completed six arrangement trials of 25 trees, with trees sampled from the whole 5x5 grid of branchiness and leafiness on each trial. At the beginning of each trial, the trees were randomly arranged in an attempt to minimise other sources of bias. The allocation of exemplars to trials was randomised across subjects. We correlated the dissimilarity matrices derived from the arrangements with a model matrix that represented a perfect grid-like arrangement to compute a "grid score" for each participant. We planned to exclude participants who failed to reach the median grid score reported in the previous study where participants were recruited online (Flesch et al., 2018), but no participants met this criterion (**Appendix B, Fig. B.1**).

Training Session (Day2): On day 2, participants took part in an online training session in which they learned to perform the task. On each trial participants first viewed a cue indicating the context (or "garden"), which was a blue or orange rectangular frame presented for 1000ms. Next, a tree was displayed for 1500ms within the frame, together with the response contingencies ("plant" or "don't plant") which were indicated by left and right arrow buttons on either side of the tree stimulus. These contingencies

(i.e. whether "plant" was mapped onto the left or right button) were varied randomly from trial to trial. The stimulus and response interval were always set to 1500ms. A response provided within this interval was highlighted by a rectangle drawn around the chosen option ("plant" vs "don't plant"). Participants were asked to learn to plant trees that grew successfully. Tree growth success depended on leafiness in one context and branchiness in another and was signalled by a numerical reward, ranging in five steps from -50 to +50. For example, for a given participant, trees occurring within the orange frame might grow successfully if they had fewer leaves, whereas trees occurring within the blue frame might grow successfully if they had more branches. Feedback, where available (see below) was presented for a period of 500ms (800ms for missed trials) and consisted of a numerical reward (if the tree grew successfully) or penalty (if it did not) for planting a tree, and always a reward of zero for not planting a tree. At the beginning of the feedback period, the tree stimulus was replaced by a fixation cross and the response contingencies were replaced by numeral rewards. These rewards/penalties were mapped onto the relevant dimension (branchiness/leafiness) and hence varied in five discrete steps from -50 to +50. Rewards (values above 0) were displayed in green, whereas penalties (rewards below zero) were displayed in red. Rewards of zero were displayed in black. Again, the chosen option was highlighted by a rectangle, with its colour matching the colour of the reward value (red/green/black). For training sessions, the intertrial interval (ITI) had a duration of 1000ms. The directionality of the rewards (more vs less leafy/branchy trees grow better) and the task order during the main training phase were fully counterbalanced across participants.

The training session consisted of three different blocks in which contexts could be either blocked or interleaved. Blocked means that all trials of one context were presented first, followed by all trials in another context, with the order counterbalanced over participants. Interleaved means that trials were shuffled so that they occurred in random order, but with exactly the same number in each context. Participants underwent a brief interleaved familiarisation phase with feedback (50 trials), followed by an interleaved baseline test (200 trials, no feedback). There was then a long main training session which was blocked (900 trials) (**Appendix B, Fig. B.1**). The purpose of the

baseline training and test was to familiarise the subjects with the task and to assess the effectiveness of the main training.

Scanning Session (Day3): The test session consisted of a brief refresher phase (interleaved, 50 trials, feedback) and the main test phase (interleaved, 600 trials, no feedback). The refresher was completed on the experimenter's laptop and was identical to the baseline training on day 2. For the test phase inside the scanner, we used a jittered ITI of 2000-6000ms (uniform) during which only the grey background was displayed. The total length of all ITIs was restricted such that all runs had equal length.

4.4.3 Data Acquisition

fMRI acquisition: Magnetic resonance images were recorded with a 3T Siemens scanner with a 32-channel head coil. A high-resolution T1-weighted structural image (voxel size = 1x1x1 mm, 176x256x256 grid, TR=1900ms, TE=2.52ms, TI=900ms) was acquired for each participant prior to the task. Each fMRI image contained 32 axial echo-planar images (EPI) in descending sequence (3.5x3.5x3.5mm isotropic, slice spacing 4.2mm, TR=2000ms, flip angle = 80, TE = 30ms). We collected fMRI data in six independent runs of 345volumes each.

fMRI Pre-processing: Pre-processing was conducted in MATLAB with SPM12 and custom scripts. For each participant, functional scans were first realigned to the first scan. As EPIs were acquired in descending sequence, we applied a slice time acquisition correction with the middle slice (TR/2=1s) as reference. Next, the structural scan was co-registered to the mean EPI. Anatomical scans were normalised to standard Montreal Neurological Institute (MNI) 152 template. EPIs were normalised to the template using tissue probability maps for grey matter, white matter, and cerebrospinal fluid. The EPIs were resliced to 3x3x3mm resolution. For univariate analyses, we applied smoothing with a full width half maximum (FWHM) Gaussian kernel of 8mm.

4.4.4 Quantification and statistical analysis

Human behavioural and fMRI data

Sigmoid fits: To estimate sensitivity of choices made by the human participants to the relevant and irrelevant feature dimensions, we fit sigmoidal curves at the level

of individual participants. First, responses were averaged across test trials and tasks within each of the five bins along a given dimension. Next, we fit a sigmoidal curve of the following form to the data, using the $curve_fit$ function of the SciPy package:

$$\sigma(x) = \frac{L + (1 - 2L)}{1 - exp(-k(x - x0))} \tag{4.1}$$

where L controlled the proportion of nonspecific errors (lapses), k the slope and $x0$ the offset of the sigmoid. Statistical inference was performed on the group-level distributions of the individually estimated parameters.

Factorised/linear model: To calculate the extent to which the participants learned a factorised solution, comprised of one accurate category boundary per task, or a linear solution, where the same boundary was applied to both tasks, we performed a model-based representational similarity analysis on the behavioural responses. First, we created choice matrices (see above) for each network run / at single subject level. We then constructed two model choice matrices, the factorised and the linear model. In the factorised model, all entries corresponding to rewarding trials were set to 1, and entries corresponding to non-rewarding trials were set to zero. Category-boundary trials were set to 0.5. In the linear model, we assumed a diagonal category boundary distinguishing between trials that were rewarding/non-rewarding irrespective of context and set the corresponding entries in the two matrices to 1, 0.5 and 0 respectively. We then concatenated the flattened choice matrices for the first and second task and constructed RDMs from the resulting vectors using the squareform and pdist functions from the SciPy package. The empirical RDMs, constructed from the behavioural responses were then regressed against the two model RDMs at the level of individual participants.

Psychophysical model: To decompose errors made by the participants into different sources, we fit a psychophysical model with five free parameters to choices at single subject level. The model had parameters for the angles of the decision boundaries in the two-dimensional stimulus space, as well as the slope, offset and lapse-rate of a sigmoidal transducer. The model projected the 2D stimulus space onto an axis per-

pendicular to the decision boundary and fed the projected values through a sigmoid to generate choice probabilities. Let X_a and X_b be the 25x2 matrices of coordinates for the stimuli of the first and second task, where each row corresponds to the x- and y-location of the peak of a Gaussian "blob". The first two free parameters θ_a and θ_b determined the angle of the line onto which these stimuli were projected:

$$X^{proj} = \begin{bmatrix} X_a \left[cos\left(\theta_a\right),\ sin\left(\theta_a\right)\right]^T \\ X_b \left[cos\left(\theta_b\right),\ sin\left(\theta_b\right)\right]^T \end{bmatrix} \tag{4.2}$$

Next, the projected values were passed through a sigmoidal transducer with free parameters for the lapse rate L, the slope k and the offset $x0$:

$$\hat{y}\left(X^{proj}\right) = \frac{L + (1 - 2L)}{1 - exp\left(-k\left(X^{proj} - x0\right)\right)} \tag{4.3}$$

We fit this model to empirical data by minimising the following loss function that quantified the mismatch between the model's output and the choices made by the participant:

$$J\left(y, \hat{y}; \theta, L, k, x0\right) = -\sum_i log\left(1 - \left|y_i - \hat{y}_i\left(\theta, L, k, x0\right)\right|\right) \tag{4.4}$$

Minimisation was performed with the L-BFGS algorithm as implemented in the minimise function of the SciPy package, with constraints set on the range of parameter values so that angle θ in $[0, 359]$, slope in $[0, 20]$, offset in $[-1, 1]$ and lapse in $[0, 0.5]$.

Group level inference: Inference was performed via paired t-tests or signed-rank tests when valuations of the assumptions of t-tests were observed.

fMRI Data Analysis: GLMs. Data were analysed using SPM12, the RSA toolbox (Nili et al., 2014) and custom scripts written in MATLAB. We used a general linear model (GLM) approach for all univariate analyses. A 128s temporal high-pass filter was applied to remove low-frequency scanner artefacts. Temporal autocorrelation was estimated with a first-order autoregressive model (AR-1). All GLMs contained regressors coding for onset and duration (boxcar until participant response) of events, which were convolved with the canonical haemodynamic response function (HRF).

Six motion parameter estimates from the pre-processing stage were included as nuisance regressors in all GLMs. Each run was represented by a separate set of regressors in the GLM, and run number was encoded by a dummy variable. Observed fMRI data at single subject level was regressed against this design matrix. Our analyses are based on four different GLMs. The first GLM (GLM1) had two predictors of interest (task switch trials and task stay trials), locked to cue onset. GLM2 included two parametric regressors of absolute distance of stimuli to the category boundary, for the relevant and irrelevant dimension, respectively. GLM3 included parametric regressors of the stimulus value and their interaction with choice. GLM4 was constructed for representational similarity analysis (RSA) and fitted to unsmoothed EPIs. It had 50 regressors per run, one for each combination of context ("north garden"/blue rectangle vs "south garden"/orange rectangle), branchiness (1 to 5) and leafiness (1 to 5).

Representational Similarity Analysis of human fMRI. GLM4 (described above) was fit to neural data at single-voxel level. We then constructed neural Representational Dissimilarity Matrices (RDMs) using a spherical searchlight (radius 12mm). For each searchlight sphere, we computed cross-validated neural RDMs from the condition-by-voxel matrix of estimated neural responses using Pearson correlation distance between pairs of conditions from distinct runs:

$$d(x_i, x_j) \;=\; 1 \;-\; \frac{(x_i - \bar{x}_i)\left(x_j - \bar{x}_j\right)^T}{\sqrt{(x_i - \bar{x}_i)(x_i - \bar{x}_i)^T (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T}} \tag{4.5}$$

This yielded a 300x300 RDM (50 conditions per run, six runs). All analyses excluded within-run similarity data (e.g. blocks of 50 conditions on the major diagonal). We constructed seven model RDMs to probe for the existence of task-related representational geometries in the fMRI activity patterns: the (1) grid model, (2) orthogonal manifold model, (3) parallel manifold model and (4) rotated grid model, (5) only branchiness model, (6) only leafiness model and (7) diagonal model. Let the vectors of branchiness and leafiness be $b = [-2,-1,0,1,2]^T$ and $l = [-2,-1,0,1,2]^T$. Let the task vector be defined as $t = [0,1]^T$. Let the matrix of all possible ordered tuples of

context, branchiness and leafiness be:

$$X^{50x3} = \{(x,y,z) : x \in t,\ \ y \in b\ \ and\ z \in l\ \ \} \tag{4.6}$$

The first three models were identical to the grid, orthogonal and parallel model described in the methods section of the previous chapter.

For the fourth model, we rotated one of the grids from the grid model by 90 degrees, considering the reward assignment the participant had been trained on (hence discriminating "plantiness" of trees, i.e. the extent to which "plant" was the correct answer). For example, if higher feature values led to larger rewards in both contexts, the rotation matrix would be defined as:

$$R_A(90) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(90) & sin(90) \\ 0 & -sin(90) & cos(90) \end{bmatrix} \tag{4.7}$$

We applied this rotation matrix to the grid from one task by stacking $X_A R_A$ and X_B:

$$X_{rotgrid} = \begin{bmatrix} X_A R_A \\ X_B \end{bmatrix} \tag{4.8}$$

The fifth and sixth models served as controls, based on the assumption that early visual areas might exhibit task-agnostic shape (branchiness) or colour (leafiness) sensitivity. These models were obtained by projecting both task-specific submatrices either onto the branchiness ($X_A Y_A$ and $X_B Y_A$) or onto the leafiness dimension ($X_A Y_B$ and $X_B Y_B$).

$$X_{branch} = \begin{bmatrix} X_A Y_A \\ X_B Y_A \end{bmatrix} \tag{4.9}$$

$$X_{leaf} = \begin{bmatrix} X_A Y_B \\ X_B Y_B \end{bmatrix} \tag{4.10}$$

The last model was obtained by taking the grid of all combinations of branchiness and

leafiness and projecting trees onto the main diagonal, ranging from low leafiness and low branchiness to high leafiness/branchiness, with the projection XP where:

$$P = \begin{bmatrix} 1 & 0 \\ 0 & cos(45) \\ 0 & sin(45) \end{bmatrix} \tag{4.11}$$

so that:

$$X_{diag} = X_{grid}P \tag{4.12}$$

This last model was based on the competing hypothesis that humans may have ignored context and optimised for a strategy that yielded 70% correct on both tasks (Flesch et al., 2018). Within a given structural ROI or searchlight sphere, we repeated the 50x50 model RDMs over runs to match the 300x300 neural RDMs, setting the within-run-dissimilarities to NaN. We then regressed z-scored vectorised neural RDMs against z-scored sets of vectorised model RDMs using a multiple linear regression at single subject level:

$$\begin{aligned} RDM_{brain} = \beta_0 + \beta_1 RDM_{grid} + \beta_2 RDM_{rotatedgrid} + \beta_3 RDM_{orth} + \beta_4 RDM_{parallel} \\ + \beta_5 RDM_{branch} + \beta_6 RDM_{leaf} + \beta_7 RDM_{diag} \end{aligned} \tag{4.13}$$

Statistical inference was performed with a group-level t-test of the regression weights against zero. Correction for multiple comparisons was conducted via non-parametric cluster correction as implemented in the SNPM toolbox (FDR threshold < 0.05). To avoid circular inference, all post-hoc visualisations and analyses within ROIs were performed in leave-one-subject-out cross-validated ROIs derived from the activity peaks identified with the searchlight approach (12 mm radius).

fMRI RSA: Parameterised Model. To obtain more fine-grained estimates of the neural geometry, we also fit a parameterised model to the cross-validated ROIs identified with the searchlight approach. As this model was identical to the one described in the methods section of the previous chapter, a detailed description is omitted here. Estimates from the model were used to visualise the representational geometries of the best

fitting RDMs via projection into three dimensions with classical Multi-Dimensional Scaling (MDS).

fMRI RSA: Embedding dimensionality. We performed Singular Value Decomposition (SVD) on the patterns of BOLD activity across voxels within each cross-validated ROI and calculated the cumulative explained variance based on the squared singular values to obtain an estimate of the embedding dimensionality (Jazayeri & Ostojic, 2021) of the neural activity patterns. To test whether the directions of largest variance were aligned with the task-diagnostic dimensions of context, branchiness and leafiness, we repeated the regression-based RSA within each cross-validated candidate region after successively removing components, starting with the smallest one. This truncated SVD allowed us to identify the minimal number of components required to successfully decode a factorised representation from the neural data.

fMRI RSA: Correlations between brain and behaviour. We performed a correlation analysis (Kendall's τ_a) to quantify the extent to which orthogonal representations at the neural level predicted accurate, axis-aligned behavioural responses. We analysed human choice patterns by computing behavioural data RDMs from the probabilities of responding "plant" to trees in each condition, i.e. as a function of each stimulus' distance to bound along the irrelevant and relevant dimension in each context. Building on previous work (Flesch et al., 2018) we fit two model RDMs to human choice patterns, called the factorised and linear models. In the factorised model, choices were aligned with the ground-truth boundaries, whereas in the linear model, a "diagonal" boundary was applied to both contexts, corresponding to the single linear boundary that optimised for accuracy whilst ignoring the context (yielding 70% correct). Fitting the factorised model to behaviour yielded an "axis-alignment score", indicating whether the participant's decision boundaries were aligned with the ground truth. We tested at the group level whether the extent to which neural geometries could be explained by the orthogonal model (neural factorisation score) significantly covaried with the extent to which the factorised model explained human choices (axis alignment score).

fMRI RSA: Comparison of patterns across regions. To test statistically whether patterns differed between EVC and DLPFC/PPC/MCC, we performed random effects

Bayesian model selection using the VBA toolbox for MATLAB. We created three regression models, consisting of (1) the branchiness and orthogonal RDM, (2) only the branchiness RDM and (3) only the orthogonal RDM. These were fit to the neural RDMs in EVC, DLPFC, PPC and MCC. We then approximated the log model evidences with the subject-specific negative BIC scores derived from the individual regression model fits. With these estimates, we performed random effects Bayesian model selection (RFX-BMS) to obtain exceedance probabilities – the probability that one model explains the data better than its competitors- and estimated model frequencies – the proportion of subjects explained by each model. We report protected exceedance probabilities, which correct the exceedance probabilities to reduce the possibility that an effect is observed due to chance.

fMRI MVPA: Decoding of relevant and irrelevant dimensions. We performed a decoding analysis to assess whether a classifier trained on the relevant dimension in one task could decode the same dimension in the other task where it was irrelevant. In theory, this should only be possible in EVC, which represents both feature dimensions irrespective of context, but not in our fronto-parietal areas of interest where irrelevant dimensions were (partially) supressed. We first obtained single-trial estimates from a whole-brain GLM estimated on the neural data. We trained a linear support vector machine (SVM) with binary outputs at single subject level with leave-one-run-out cross validation on the t-maps obtained from this GLM. Within each run, patterns were first standardised and denoised by removing all but the first n principal components required to explain 95% of variance. Then, the classifier was trained to predict the choice-diagnostic label of the relevant dimension (for example, not leafy vs leafy) and its accuracy was assessed on data from the held-out run. We tested whether it could predict the relevant dimension of the task it had been trained on, the irrelevant dimension of the same task, and the relevant and irrelevant dimensions of the task it had not been trained on. Decoding accuracies were averaged within subjects across all held-out test sets and across tasks. To assess significance of the decoding performance, we performed group-level t-test against a chance level of 0.5 across subjects.

Non-human primate data

Representational Similarity Analysis of NHP electrode recordings. We created pseudo-populations by concatenating all recorded units, separately for monkey A and monkey F. Unit-by-stimulus response matrices were obtained by averaging activity across trials for each stimulus type (6 motion directions * 6 colours * 2 contexts = 72 entries). RDMs were constructed from these matrices using the Euclidean distance measure. For all reported analyses, we focus on activity averaged over the second half of the trial (300-600ms) where task factorisation was strongest, an observation consistent with previous reports of dynamic encoding of different task variables throughout a trial (Aoi et al., 2019). We fitted the same set of candidate model RDMs to this dataset as previously to RDMs obtained from human fMRI data (see above). For statistical inference, we created a null distribution by randomly permuting the trial labels and repeating this regression-based RSA 1000 times. We calculated p-values from the proportion of permutations that yielded regression coefficients larger than the one observed on the original data.

Individual unit selectivity and axis alignment of NHP electrode recordings. We assessed task selectivity of individual units using a standard regression-based approach. Mean activity of each unit was regressed against four predictors, coding for colour and motion direction separately for each context:

$$y_{unit} = \beta_0 + \beta_1 \,\text{colour}_{colour\ task} + \beta_2 \,\text{motion}_{colour\ task} + \beta_3 \,\text{colour}_{motion\ task} + \beta_4 \text{motion}_{motion\ task} \tag{4.14}$$

Selectivity was defined as having a significant regression coefficient for the variable of interest. Due to the substantial number of tests, we performed FDR correction to correct for multiple comparisons. We distinguish between diverse types of selectivity. Task-selectivity was defined as having a significant regression weight only for the relevant feature dimension (i.e., only for motion in the motion task or colour in the colour task). Task-agnosticity was defined as having significant coefficients for both dimensions. Furthermore, we identified units that were selective only to colour or motion, irrespective of context, and defined non-specific selectivity as having sig-

nificant regression weights that do not fall into any of the above categories. As for the hidden units in the neural network, we again plotted the different proportions of selectivity patterns of units within a pseudo population and visualised the response profile of task and stimulus selective units by averaging the activity within a sub-population separately for each combination of feature values (colour, motion) and context. Axis alignment of these response matrices was assessed by regressing them against the factorised and diagonal model as previously described for the neural network (see above). We assessed the embedding dimensionality of the patterns observed in monkey FEF using the same truncated SVD approach described above for the human fMRI data.

Decoding of relevant and irrelevant dimensions. To assess whether task-irrelevant dimensions were filtered out in the NHP data, we performed a decoding analysis that was similar to the one described above for the fMRI data. We first divided the trials of each unit into a first and second half. As data from the units in the original dataset had been recorded during different sessions, we first created fully counterbalanced pseudo trials. We generated 1440 pseudo trials by sampling each condition from the set of recording units and creating vectors of condition-by-unit activity that represented individual trials as if activity from these units had been recorded simultaneously. We repeated the procedure for the second half of the dataset, thus yielding 1440 training and 1440 test trials. The data was standardised and denoised by removing all but the first n principal components required to explain 90% of the variance. We then trained a linear SVM on the relevant dimension of the NHP dataset with two-fold cross-validation and assessed its decoding performance on the relevant and irrelevant dimensions in the held-out dataset. Statistical significance was assessed with a permutation test in which we computed test performance after randomly shuffling the labels (1000 permutations). Chance was defined as the average performance on these shuffled datasets (roughly 0.5%) and p-values were computed from the proportion of trials in which the decoding accuracy exceeded the one observed on the original data.

Chapter 5

A neural network model of human continual learning

Abstract

Humans can learn several tasks in succession with minimal mutual interference but perform more poorly when trained on multiple tasks at once. The opposite is true for standard deep neural networks. Here, we propose novel computational constraints for artificial neural networks, inspired by earlier work on gating in the primate prefrontal cortex, that capture the cost of interleaved training and allow the network to learn two tasks in sequence without forgetting. We augment standard stochastic gradient descent with two algorithmic motifs, so-called "sluggish" task units and a Hebbian training step that strengthens connections between task units and hidden units that encode task-relevant information. We found that the "sluggish" units introduce a switch-cost during training, which biases representations under interleaved training towards a joint representation that ignores the contextual cue, while the Hebbian step promotes the formation of a gating scheme from task units to the hidden layer that produces orthogonal representations which are perfectly guarded against interference. Validating the model on previously published human behavioural data revealed that it matches performance of participants who had been trained on blocked or interleaved curricula, and that these performance differences were driven by misestimation of the true category boundary.

5.1 Introduction

Humans have the remarkable ability to learn multiple tasks over their lifespan. New tasks can be learned in sequence with minimal disruption to previously acquired tasks, a feat that is known as continual learning. For example, in the case of (supervised) visual categorisation, if so asked you could learn to successfully categorise fruits by size (crab apple vs. granny smith) and then by colour (ripe vs. unripe) without the latter learning overwriting the former. Building neural networks that learn continually has proved challenging in AI research (Hadsell et al., 2020; Parisi et al., 2019). In neuroscience, it remains an open question how the human brain learns continually, and whether biology can inspire candidate solutions for artificial agents (Carvalho & Goldstone, 2014, 2015; Franklin & Frank, 2020; Musslick & Cohen, 2021; Wulf & Shea, 2002). Here, we present a computational model of human continual learning, which builds on earlier work presented in the previous two chapters.

With the advent of deep learning, artificial neural networks are enjoying a renaissance as models of biological information processing (Richards et al., 2019; Saxe et al., 2021). Despite their architectural simplicity, representations that emerge in neural networks bear striking similarities to those observed in early visual cortex and higher association areas of the human brain (Güçlü & Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Lindsay, 2021; Yamins et al., 2014; Zhuang et al., 2021), leading to the proposal that these models can be used as a test-bed for theories about the geometry (Flesch et al., 2022; Yang et al., 2019) and dimensionality of neural representations (Badre et al., 2021; Ito & Murray, 2021), or the feature selectivity of downstream cortical areas (Jagadeesh & Gardner, 2022). However, without significant modification, neural networks trained with gradient descent are unable to learn multiple tasks in sequence without suffering from catastrophic forms of forgetting (French, 1999; Lee et al., 2021; Parisi et al., 2019). Accordingly, the problem of catastrophic forgetting has received considerable attention in the machine learning community, and numerous engineering solutions have been proposed that wholly or partially prevent forgetting, either by preventing task-relevant weights from changing (Kirkpatrick et al., 2017; Zenke et al., 2017), dynamic architecture growth (Rusu et al., 2016), (deep

generative) replay (Shin et al., 2017) or orthogonalization of representations in the hidden layer (Chaudhry et al., 2020; Farajtabar et al., 2019; Zeng et al., 2019). Some solutions draw loose inspiration from neuroscience, such as experience replay, which is motivated by Complementary Learning Systems theory (McClelland et al., 1995; van de Ven et al., 2020) or gating approaches, linked to top-down attentional control (Masse et al., 2018), or regularisation approaches, which can be related to changes in synaptic plasticity on different timescales (Benna & Fusi, 2016; Kaplanis et al., 2018), it remains unclear which of these methods can best capture the ways in which humans learn and execute multiple tasks in series.

In neuroscience, the term "cognitive control" is applied to neural mechanisms that allow a context-appropriate task to be selected and executed. Cognitive control has long been associated with the prefrontal cortex (PFC), based on evidence that prefrontal neurons code for specific tasks, and exert top-down control to prioritise context-appropriate stimuli and actions (Buchsbaum et al., 2005; Johnston et al., 2007; Mansouri et al., 2006; Miller & Cohen, 2001; Rikhye et al., 2018; Rougier et al., 2005). In the domain of categorisation, it has been proposed that the PFC may implement cognitive control by gating (or compressing) task-irrelevant input dimensions (Cohen et al., 1990; Gisiger & Boukadoum, 2011; Tsuda et al., 2020). More recent investigations of neural geometry have shown that during multi-task performance, mutual interference among tasks is minimised by projecting relevant dimensions into orthogonal, low-dimensional subspaces (Flesch et al., 2022; Ito & Murray, 2021; Libby & Buschman, 2021; Panichello & Buschman, 2021). A key challenge, thus, is to identify a mechanism that can learn from scratch to protect sequentially occurring task representations from mutual interference.

We sought to develop a biologically inspired neural network model that describes how humans learn to perform multiple tasks in series. A starting point for our work is the observation that humans actively benefit when categorisation tasks are temporally autocorrelated (blocked) during training. For example, consider a validation task which requires naturalistic stimuli (tree images) to be categorised alternately by dimensions of leaf and branch density. Humans benefit from a training regime con-

sisting of long training blocks of unidimensional leafy or branchy rules, rather than training blocks in which leafy and branchy rules are interleaved together (Flesch et al., 2018). This benefit appears to be particularly pronounced when exemplars are highly heterogenous within and across tasks (Carvalho & Goldstone, 2014, 2015). Thus, our goal was to identify a model that could learn from scratch to capture the benefit of blocking and the cost of interleaving, as well as the patterns of neural geometry that have been observed during multitask performance.

There are two key ideas that motivate our model design. The first is that biological neural circuits have intrinsic time constants of integration which ensure that decisions are driven by information from the immediate past as well as the present. This principle underlies ubiquitously observed trial history effects in decision tasks (Cho et al., 2002; Soetens et al., 1985; Yu & Cohen, 2008). The second is that simple learning based on coincidence detection (such as Hebbian learning) allow groupings of inputs to be effectively orthogonalised. Our model capitalises on these principles by combining two algorithmic motifs. Firstly, we assume that neuronal responses are "sluggish": on each trial, inputs to the network contain some information carried over from previous trials. Carrying over contextual cues from previous trials increases task interference (switch costs) in interleaved conditions (where sequential trials may require performance of conflicting tasks) but not in blocked conditions (where sequential trials involve the same task). Secondly, we propose that a Hebbian learning step follows each supervised parameter update, to strengthen connections between task signalling units and hidden units that encode task-relevant information. This has the effect of orthogonalising the weights linking context to hidden units for the two tasks, allowing tasks to be represented in independent subspaces in the hidden layer (Flesch et al., 2021). This intervention thus implements a form of context-dependent gating (Cohen et al., 1990; Gisiger & Boukadoum, 2011). However, in contrast to earlier work on cognitive control and related papers that have used gating as a means for continual learning (Masse et al., 2018; Russin et al., 2022; Tsuda et al., 2020), we demonstrate that this control signal can be acquired by a simple biologically-inspired mechanism and without direct intervention by the experimenter. Finally, we show that this model forms highly

task-specific neural codes, similar to those reported in a series of recent studies on the geometry of representations in human and macaque prefrontal cortex (Flesch et al., 2022; Ito & Murray, 2021; Libby & Buschman, 2021; Panichello & Buschman, 2021)

5.2 Results

All simulations described here involve the binary categorisation of stimuli according to one of two task rules, which are defined by orthogonal category boundaries in feature space. In all simulations, the rules are explicitly cued by a contextual signal, and fully supervised feedback is provided based on the context (task) and stimuli (Mante et al., 2013). Thus, one can conceive the model as performing a task in which trees are categorised by leaf and branch density, or apples by size and colour. In practice, network inputs were simplified images of Gaussian "blobs", with the two relevant dimensions being the location of the peak along the x-and y-axis respectively (**Fig. 5.1A**). This allowed us a testbed that matched our domain of interest (e.g., inputs were high-dimensional, but two cardinal dimensions were relevant) without the potential biases that arise from naturalistic stimuli. We refer to the two "tasks" performed by the neural network as discriminating the peak of the blob with respect to lines that bisected the horizontal and vertical midlines respectively (**Fig. 5.1A**).

5.2.1 Blocked vs interleaved training with standard SGD

We began by evaluating a model we call the "vanilla SGD" network. The model is a fully connected feedforward network (multi-layer perceptron or MLP) with a single hidden layer, Rectified Linear Unit (ReLU) non-linearities and a single output node. We initialized the network with small random weights ($\sigma = 0.001$), placing the network in the feature learning regime (Flesch et al., 2022). Inputs to the network were flattened images of Gaussian blobs, together with a one-hot encoded contextual cue signal (e.g. $[0,1]$ for task 1 and $[1,0]$ for task 2; see **Fig. 5.1B**. The network was trained using stochastic gradient descent (SGD) either on blocked data, where it was exposed to one task at a time over a prolonged training block, or on interleaved data where trials from both tasks were randomly interspersed within a single block (**Fig. 5.1C**). It was then evaluated on both tasks without supervision (i.e., with no further

optimisation).

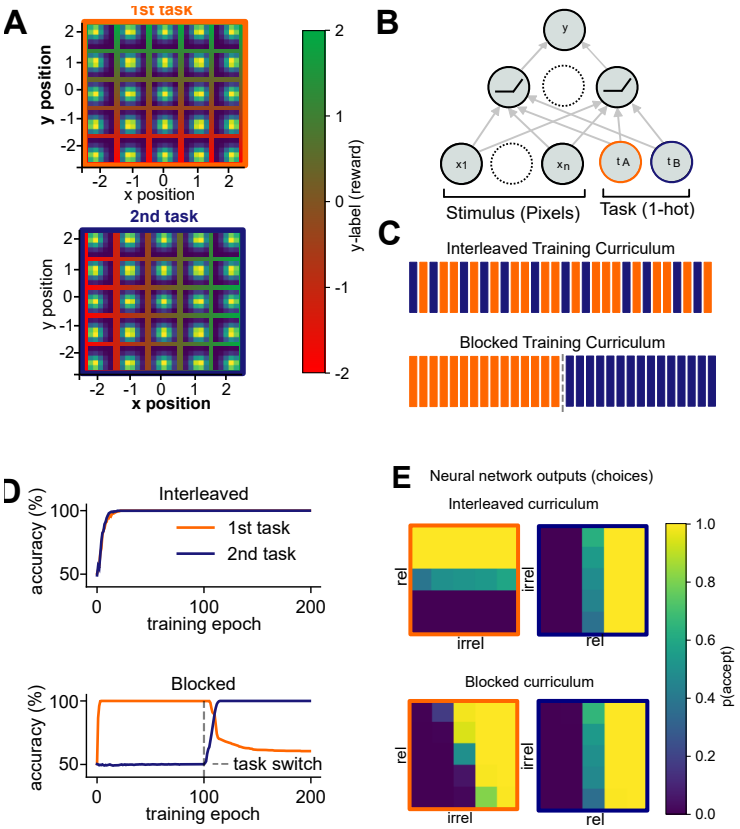

Figure 5.1: Task design, network architecture and results of blocked versus interleaved learning. (**A**) Task design. Stimuli were two-dimensional Gaussian functions ("blobs") for which we systematically varied the location of its peak along the x- and y- dimensions in five discrete steps. Each subpanel visualises the Gaussian blob input image at that location in the underlying 2D stimulus space. Only one of the two feature dimensions was relevant per task, so that the reward (y-label) depended on the x-position in the first task (orange) and y-position in the second task (blue). (**B**) The network was a simple feed-forward MLP with a single hidden layer with ReLU non-linearities and received the flattened images of Gaussian blobs together with a one-hot encoded task signal as inputs. (**C**) The network was trained either in a fully interleaved curriculum in which trials from both contexts were randomly interspersed, or in a blocked curriculum in which it was first trained on one task, and then on the other. (**D**) Under interleaved training, the network quickly reached 100% training accuracy on both tasks. In contrast, under blocked training, learning the second task came at the cost of forgetting how to perform the first task. (**E**) Plotting the choices of the trained network in two dimensions revealed that under interleaved training, choices were aligned with the ground truth category boundaries (shown in (A)), whereas under blocked training, the network treated the first task as if it was the second.

As expected, the vanilla SGD network converged to perfect performance under interleaved training but suffered catastrophic interference when trained on each task

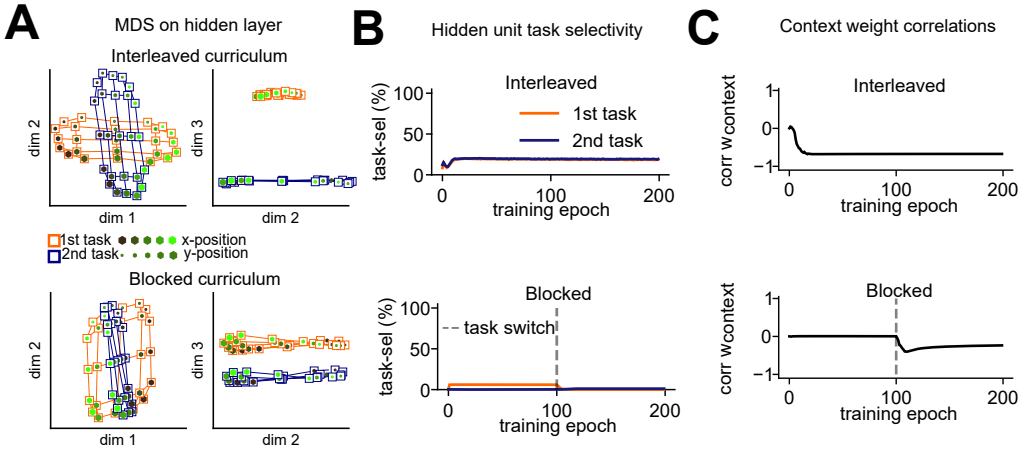

Figure 5.2: Hidden layer representations under blocked and interleaved learning. (**A**) Projections of the hidden layer activity into three dimensions via multi-dimensional scaling (MDS) shows orthogonal representations under interleaved training, where irrelevant information was suppressed, and parallel representations under blocked training, where the first task is encoded in the same way as the second task. (**B**) Under interleaved training, a significant proportion of hidden units were exclusively selective to the relevant dimension in one task (but not the other), whereas no such task-selectivity was observed under blocked training. (**C**) Evolution of correlation between context weights for both tasks during training. Interleaved - but not blocked - training promoted learning of anti-correlated context weights.

in succession, with the ability to perform the first task overwritten by training on the second (**Fig. 5.1D**). Plotting network choices made during validation as a function of the two feature values (x- and y-location) revealed that under interleaved training, the network learned two orthogonal category boundaries, one per task. Under blocked training, however, it applied the category boundary of the second task to the first task, ignoring the task signal (**Fig. 5.1E**). Projecting the hidden layer representations observed during validation into two dimensions confirmed that this network had learned task-specific manifolds under interleaved training. Each task was represented by a single axis that only encodes task-relevant information – the location along the x- or y-axis respectively. The axes were orthogonal to each other and separated by context along the third direction (**Fig. 5.2A**, upper), a finding we had already observed in a previous study (Flesch et al., 2022). In contrast, after blocked training, the network represented the first task as if it were the second, and no longer distinguished between tasks (**Fig. 5.2A**, lower).

How did the network learn this representation? Previous work suggested that the pattern observed under interleaved training can be obtained via non-linear gating, if the context signal acts as additive bias to filter out irrelevant dimensions via context-dependent deactivation of units that encode task-irrelevant information (Flesch et al., 2022). In fact, 40% of units in the hidden layer became task-selective under interleaved training, responding to the relevant (but not irrelevant) dimension in one task and being active in the other task (**Fig. 5.2B**, upper). Under blocked learning, however, no such task-selective units emerged, suggesting that the network ignored the task signal (**Fig. 5.2B**, lower). We have previously observed that the weights from the task units to the hidden units become anti-correlated over the course of interleaved training, pushing the input to the ReLU to positive or negative values depending on the context (Flesch et al., 2022). For the current simulations, this effect is shown in **Fig. 5.2C**. Under blocked learning, this anti-correlation does not emerge, as the network fails to utilise the task signal (**Fig. 5.2C**). Taken together, thus, we found that in the vanilla SGD network, the two tasks were represented by allocating them independent hidden layer units, using context-dependent gating. This replicates our earlier report (Flesch et al., 2022). Under blocked training, the network failed to utilise the task units to implement this gating scheme, as the task signal was not required to solve individual tasks in isolation.

5.2.2 Modelling the cost of interleaving with "sluggish" neurons

During validation, humans are less accurate after interleaved compared to blocked training on the visual categorisation task (Flesch et al., 2018). In other words, they seem to show opposite behaviour to the vanilla SGD network, which had lower performance on blocked compared to interleaved training. We thus sought to develop a theory that could account for these discrepancies and devise algorithmic motifs that would more closely mimic those performance differences observed in human participants. How does this cost of interleaved training arise? In the real world, contexts tend to be temporally autocorrelated. Humans spend prolonged periods of time in one context, and switches occur intermittently (for example, when you leave the office to head home for the day, or when you leave the motorway and drive through an urban

area). One possibility, thus, is that participants have an inductive bias that tasks should remain the same over time, in which case it is rational to condition behaviour not just on current task cues, but those that occurred in the immediate past (Yu & Cohen, 2008). This explanation has been offered for the ubiquitous observation that people are biased by the cues and responses that occurred on previous trials, and that switching between tasks incurs a cost to accuracy and RT (Monsell, 2003). Here, we propose that in humans, these choice history biases create interference during interleaved, but not during blocked learning (see (Russin et al., 2022) for a related account). Previously, we hypothesised that this may lead humans to ignore the context signal and effectively apply the same categorisation rule irrespective of the context, which optimises for performance on congruent trials (those with the same responses across tasks.) (**Fig. 5.3A**, lower) (Flesch et al., 2018). In contrast to this linear solution, with blocked training, human participants can effectively factorise the decision problem and learn one rule per task (**Fig. 5.3A**, upper).

To model this cost and tendency towards a linear solution, we introduce the concept of "sluggish" units, that is, neurons that carry information from previous trials over to the current trial (Flesch et al., 2021). We model this sluggishness with an exponentially moving average (EMA, see methods) with the weight on previous trials controlled by a single parameter, α. Setting $\alpha = 0$, is equivalent to the vanilla SGD network described above; other models are "sluggish SGD" networks. Increasing α has the effect of decreasing performance at validation overall (**Fig. 5.3B**). In **Fig. 5.3C**, we plot psychometric data, i.e., the effect of α on how response probability varies with relevant and irrelevant information. Visual inspection suggests that the parameter controls the extent to which information along the irrelevant dimension is factored into the model's choices (**Fig. 5.3C**).

Plotting the choices in two dimensions offers further insights into the effect of sluggishness. As α increases, the model moves from learning a factorised solution with one boundary per task to a linear solution with a single category boundary (**Fig. 5.3E**). Indeed, the factorised model fit better for low sluggishness values, whereas the linear model fit better for larger sluggishness values (**Fig. 5.3F**). In other words,

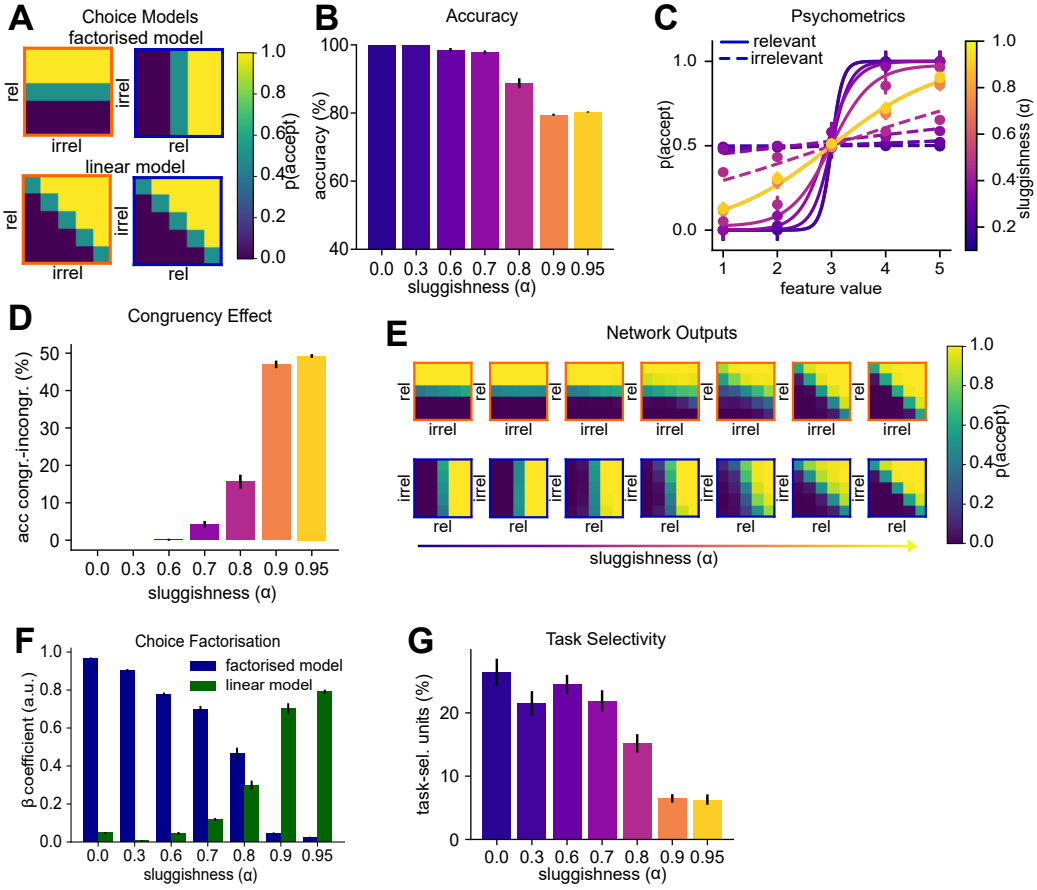

Figure 5.3: Modelling the cost of interleaving with a "sluggish" task signal. (**A**) The factorised model (top) encodes two separate category boundaries, one per task. The linear model (bottom) ignores the context signal and encodes a diagonal category boundary that yields high performance on both tasks. We hypothesised that interleaved training would promote a solution as predicted by the linear model. (**B**) Test accuracy after interleaved training with different levels of "sluggishness" (exponential average of the task signal). (**C**) Sigmoidal curves fit to the choices of networks described in (B). Solid lines for relevant dimensions, dashed lines for irrelevant feature dimensions. As the sluggishness increases, sensitivity to the relevant dimension decreases and to the irrelevant dimensions increases. (**D**) Congruency effect as a function of amount of sluggishness. (**E**) Network outputs (choices) for different levels of sluggishness. As sluggishness increases, the networks move from learning a "factorised" to learning a "linear" solution. (**F**) Coefficients obtained from regressing the outputs shown in (E) against the models shown in (A), confirming that sluggishness controls whether a factorised or linear solution is learned. (**G**) Proportion of task-selective hidden units. Decreases with amount of sluggishness.

the sluggishness introduces a congruency effect, whereby the network performs much better on trials with the same label across tasks (congruent) compared to trials with

task-unique labels (incongruent) (**Fig. 5.3D**). At the level of neural representation, we observed a reduction of the proportion of task-selective hidden layer units (with axis aligned tuning profile) relative to task-agnostic units (selective for congruent trials) (**Fig. 5.3G**).

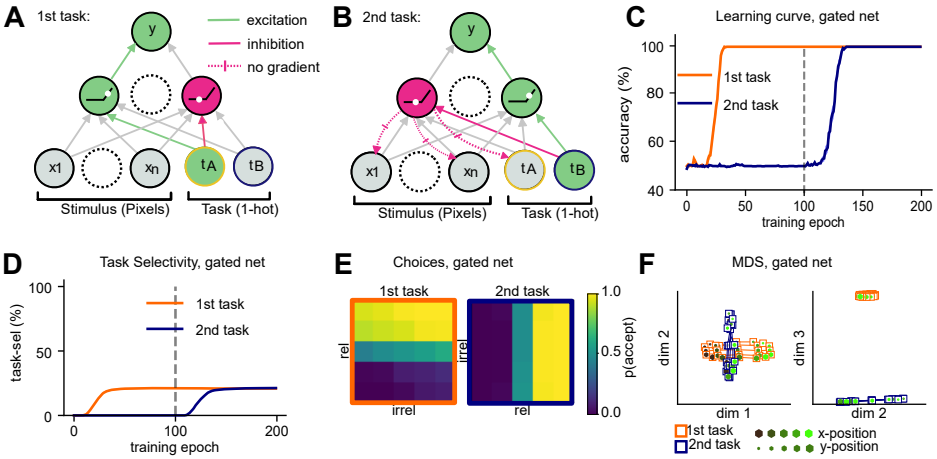

Figure 5.4: Blocked training with manual context gating. (**A**) Illustration of how task weights can be set to values with opposing signs to activate and deactivate units in the hidden layer while the network is trained on the first task. (**B**) Same as (A), but for the second task. Manually settings task weights to anti-correlated values (different signs for first and second task) ensures that different units are active in each task. (**C**) Learning curves of networks trained with manual gating, showing that forgetting has been mitigated. (**D**) As training progresses, task-selective units emerge for both tasks. Gating protects units selective for the first task during training on the second task. (**E**) Network outputs are axis-aligned, just as seen previously under interleaved training. (**F**) MDS on the hidden layer revealed orthogonal representations where task-irrelevant dimensions are attenuated.

5.2.3 Modelling blocked learning with non-linear gating

While the introduction of "sluggish" neurons imposes a cost on interleaved training, it doesn't solve the problem of catastrophic forgetting under blocked training. How can we account for the ability of humans to learn continuously without substantial forgetting? The vanilla SGD network trained on interleaved data learned a factorised representation where different populations of hidden units were allocated to the first and second task. This allocation was achieved via non-linear gating, implemented by the context weights which pushed the hidden layer activity into the negative/positive input range of the ReLU non-linearities. We wondered whether this simple gating mechanism that allocates different subsets of units to different tasks may be sufficient

to guard against forgetting. To test this, we first hand-crafted the gating scheme by manually setting the weights that connect task units to hidden units to anti-correlated values, such that each unit received a positive bias in one task and a negative bias in the other (**Fig. 5.4A,B**). We then trained the remaining units end-to-end on a blocked curriculum. This network no longer forgot how to perform the first task after it was trained on the second (**Fig. 5.4C**), which suggests that a simple gating intervention that partitions the hidden layer may be sufficient to guard against catastrophic interference. The outputs of the network were axis-aligned, demonstrating that it learned accurate representations of the two category boundaries (**Fig. 5.4E**). At the level of hidden units, we observed once again orthogonal and low-dimensional manifolds that encoded task-relevant and suppressed task-irrelevant dimensions in a context-dependent manner, just like in the vanilla SGD network trained on interleaved data (**Fig. 5.4D,F**). Note that (Russin et al., 2022) describes a closely-related set of simulations and equivalent results in this handcrafted setting.

5.2.4 Hebbian learning of anti-correlated context weights

Ideally, we would like these gating signals to be acquired without intervention by the experimenter. Thus, we introduced another algorithmic motif: the use of a Hebbian learning step following supervision. Due to the one-hot representation of the context variable, the context units are correlated with those hidden units that encode task-relevant information for the active context. Consequently, with mean-centered inputs, the Hebbian step strengthens the connections between the task context units and those hidden units encoding task-relevant information and weakens the connection to units coding for irrelevant information. We use a variant of Hebbian learning with weight-decay, called Oja's rule (Oja, 1982; Oja & Karhunen, 1985). A well-known property of Oja's rule is that it converges to the first principal component of the inputs when applied to mean-centred data. Crucially, in our simple case of only two tasks, the direction of largest variance in the mean-centred input space of our Gaussian blob dataset is spanned by the two task signals (**Fig 5.5A,B**). Indeed, when performing weight updates with Oja's rule on a single hidden unit, that unit recovered the first principal component of the input dataset, which distinguished between the two contexts. We ob-

served that the two weights between the context units and the hidden unit converged to values with opposing signs, the desired requirement for non-linear gating (**Fig. 5.5C**).

We concluded that Hebbian updates with Oja's rule could be used to establish links between the task signal units and active units in the hidden layer. To implement this, we extended this approach to multiple hidden units, so that each of these would learn to receive task signals via anti-correlated weights. As for the handcrafted solution in **Fig. 5.4**, when stimuli were propagated forward through the network to the hidden layer, those units that had positive outputs for task A had negative outputs for task B and vice versa. Thus, applying a ReLU nonlinearity partitions a portion of the hidden layer into task A and task B selective units. To assess whether this Hebbian learning step would be sufficient to guard against catastrophic forgetting, we devised a new training scheme in which we alternated the supervised SGD update and the Hebbian update on each training step (methods). We call this model the "Hebbian Gating" network. Crucially, we found that this intervention was sufficient to alleviate catastrophic forgetting. The performance of the network on the first task remained at ceiling, even after training on the second task (**Fig. 5.5D**). Just as in the vanilla SGD network trained on interleaved data, we observed that for the Hebbian Gating network the learned context weights were anti-correlated even for blocked training (**Fig. 5.5F**). Thus, the hidden layer was partitioned into task A and task B selective units (Fig. 5.5E) and the representations embedded in the hidden layer population response became orthogonal, with compression along the irrelevant dimensions (**Fig. 5.5H**), a factorisation that was also reflected in two accurate category boundaries at the output level (**Fig. 5.5G**).

To summarise, we have demonstrated how a variant of Hebbian learning can be used to learn anti-correlated weights that connect task units to relevant hidden units, and that alternating between supervised and Hebbian training updates allows a network trained on blocked data to learn tasks sequentially without forgetting. Representations formed by the network were identical to those observed under interleaved training in the vanilla SGD network.

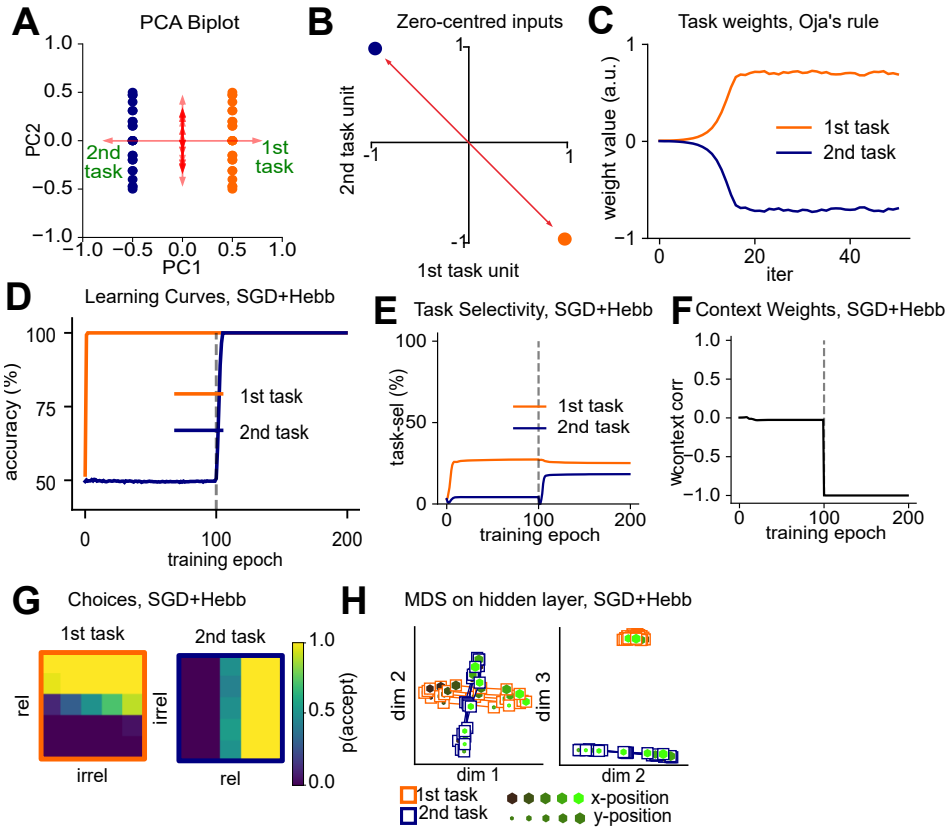

Figure 5.5: Protection against catastrophic forgetting via Hebbian learning. (**A**) Biplot on training trials. The axis of largest variation is spanned by the context signal. (**B**) Schematic showing why the first component is spanned by the context signal (for zero-centred inputs). (**C**) Weights from task units to a single hidden unit, trained with Oja's rule. Over time, weights with opposing signs emerge. (**D**) Learning curve of neural networks trained with SGD and Oja's rule in alternation. The network no longer forgets how to perform the first task. (**E**) Task selectivity of hidden layer units. SGD+Hebbian learning prevents the network from losing units that are selective to the first task. (**F**) Hebbian learning induces the desired anti-correlation between task weights. (**G**) Network outputs are axis aligned under blocked learning, just as previously seen under interleaved learning. (**H**) MDS on the hidden layer, revealing orthogonal representations that encode both tasks without interference.

5.2.5 Modelling human continual learning

Next, we assessed whether our two algorithmic innovations, the sluggishness and the Hebbian update step, were sufficient to reproduce error patterns made by human participants who had been trained on a comparable task. We re-analysed a dataset from a previous study in which participants learned to accept/reject images of fractal tree stimuli in two different task contexts, introduced as the north and south garden (Flesch

et al., 2018). Just as for our Gaussian blobs, trees varied along two different feature dimensions, corresponding to the density of leaves ("leafiness") and number of branches ("branchiness"), of which only a single dimension was relevant for each task. The participants were trained either on a blocked curriculum, or on a randomly interleaved curriculum. Crucially, participants whose training phase was blocked performed better at a subsequent interleaved validation phase, compared to those who received an interleaved training curriculum. Further analyses of the error patterns revealed that these participants had better estimates of the decision boundaries for each task and were less influenced by variation along the task-irrelevant dimensions. To assess the effectiveness of our approach, we compared validation performance after blocked or interleaved training between a neural network with both innovations, the sluggishness and the Hebbian updates (called "sluggish Hebbian gating network"), and a standard feed-forward neural network that was trained without any further algorithmic innovations ("vanilla SGD network"). To perform statistical inference on the neural networks, we collected 50 independent training runs with randomly initialised networks per training curriculum. In contrast to this baseline MLP, the network equipped with sluggishness and Hebbian update step qualitatively recreated all key aspects of the human behavioural data.

First, human participants trained on a blocked curriculum had a higher test accuracy than those trained on interleaved data ($T(93) = 2.32, p = 0.022$, **Fig. 5.6A, left panel**). While the opposite was true for the vanilla SGD network, which suffered from catastrophic interference ($T(98) = -184.33, p < 0.0001$, **Fig. 5.6A, middle panel**), the sluggish Hebbian Gating network showed a similar benefit of blocked over interleaved training at test ($T(98) = 13.11, p < 0.0001$, **Fig. 5.6A, right panel**). Our modelling of the impact of sluggishness on task performance revealed a congruency effect: The "sluggish" network performed better on congruent than incongruent trials. Hence, we wondered whether participants showed a similar congruency effect, and whether this difference would be larger in the interleaved group, where participants tended to use the same decision boundary for both tasks. Indeed, human participants showed a strong interaction between the training curriculum and the congruency ef-

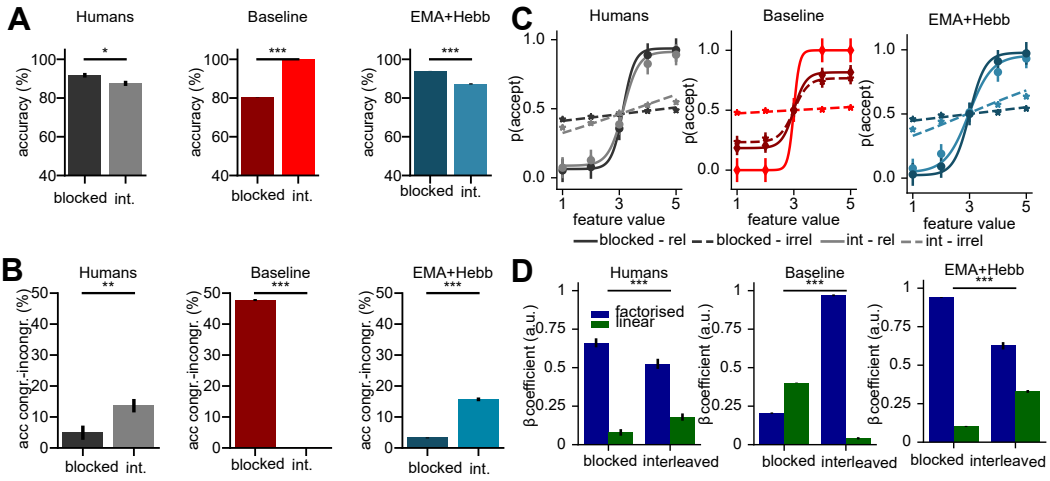

Figure 5.6: Modelling the benefit of blocked over interleaved training observed in data from human participants. (**A**) Test phase accuracy. Humans perform better under blocked compared to interleaved training, while the baseline model performs better under interleaved training, due to catastrophic forgetting. Our Hebbian network with sluggish task units performs better under blocked training. (**B**) Congruency effect. Just like human participants, under interleaved training, our model performs worse on incongruent compared to congruent trials. (**C**) Sigmoidal fits of choices made by human participants, the baseline network, and our network. Our network recreates the intrusion of irrelevant dimensions observed under interleaved, but less so under blocked training. (**D**) Fits of the factorised and linear model to human and network choices. In contrast to the baseline model, our model recreates patterns observed in humans, where the factorised model fits better under blocked than interleaved training.

fect, which was larger under interleaved training (*congruency blocked vs interleaved:* $T(93) = -2.74, p = 0.007$, **Fig. 5.6B, left panel**). Due to catastrophic forgetting, the congruency effect was much larger under blocked training in the vanilla SGD network ($T(98) = 138.09, p < 0.0001$, **Fig. 5.6B, middle panel**), while our novel training procedure for the sluggish Hebbian Gating network recreated the effect observed in humans ($T(98) = -19.30, p < 0.0001$, **Fig. 5.6B, right panel**). Next, we fitted psychometric functions (sigmoid) to the choices made by human participants and by our models, separately for the relevant and irrelevant feature dimensions. In humans, slopes for the irrelevant dimension were significantly steeper under interleaved than blocked training, suggesting that choices of these participants were stronger influenced by task-irrelevant information (*blocked vs interleaved:* $T(93) = -2.77, p = 0.0068$, **Fig. 5.6C, left panel**). Choices made by the vanilla SGD network followed the op-

posite pattern, with more intrusions from irrelevant dimensions under blocked training (*blocked vs interleaved:* $T(98) = 62.77, p < 0.0001$, **Fig. 5.6C, middle panel**). In contrast, the sluggish Hebbian Gating network was less influenced by irrelevant feature dimensions under blocked compared to interleaved training (*blocked vs interleaved:* $T(98) = -23.01, p < 0.0001$, **Fig. 5.6C, right panel**).

How did participants learn the two tasks? The original paper suggested that human participants learned "factorised" representations under blocked, but less so under interleaved training. To test this, we fit the factorised and linear model described earlier to the choices made by the models. For human participants, the factorised model explained choices better under blocked than under interleaved training ($T(93) = 3.07, p = 0.0028$, **Fig. 5.6D, left panel**), while the opposite was true for the linear model ($T(93) = -3.12, p = 0.0024$, **Fig. 5.6D, left panel**). As expected, the opposite patterns were observed for the vanilla SGD network ($T(98) = -97.96, p < 0.0001, T(98) = 46.48, p < 0.0001$, **Fig. 5.6D, middle panel**), which learned to factorise the problem under interleaved, but not blocked training. The sluggish Hebbian Gating network recreated the patterns observed in humans, suggesting that it learned two accurate decision boundaries under blocked, but not under interleaved training ($T(98) = 12.70, p < 0.0001, T(98) = -22.80, p < 0.0001$, **Fig. 5.6D, right panel**).

However, intrusions from the irrelevant dimensions might not have been the only source of errors. It was also possible that one group made more unspecific errors (lapses), was less sensitive to information along the relevant dimension or exhibited a systematic bias in the offset of their learned category boundary. Using a psychophysical model with free parameters for the angle of the learned category boundary, the number of unspecific errors, the slope and offset of the sigmoidal transducer showed that the length of training blocks predominantly affected the accuracy of the category boundary estimate (Flesch et al., 2018). Our reanalysis of the human behavioural data confirmed this, with larger angular biases in the interleaved compared to the blocked group and a significant difference in slope, while differences in

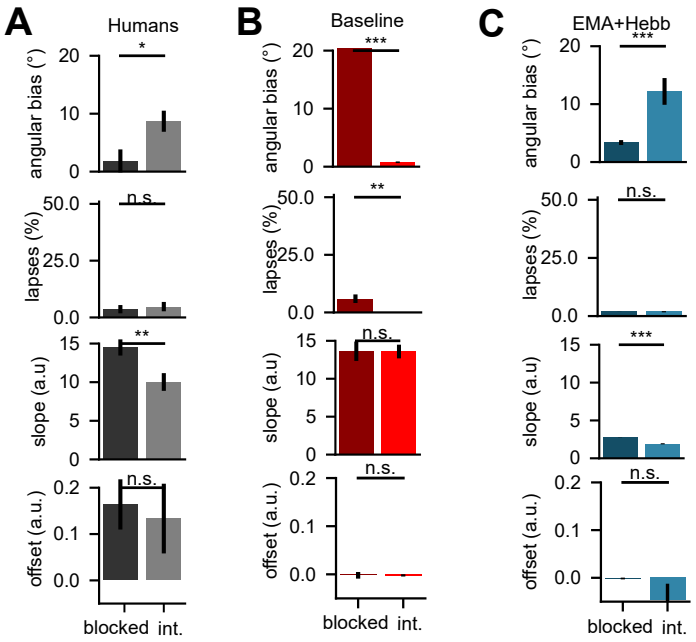

Figure 5.7: Fits of Psychometric model to human behaviour and neural networks. (**A**) Fits of a psychometric model with four parameters to choices made by human participants, decomposing error patterns into (i) angular bias of category boundary (ii) lapse rate (iii) slope and (iv) offset of sigmoidal transducer. Both the angular bias and slope differed significantly between participants trained on blocked and interleaved curricula. (**B**) Same as (A) but fitted to the vanilla SGD network. (**C**) Same as (A), but fitted to the sluggish Hebbian Gating network.

lapse and offset parameters were non-significant (*bias:* $T(93) = -2.54, p = 0.0127$, *lapse:* $T(93) = -0.41, p = 0.6807$, *slope:* $T(93) = 2.88, p = 0.0049$, *offset:* $T(93) = 0.33, p = 0.7419$, **Fig. 5.7A**). The vanilla SGD network had significantly larger angular bias and lapse rates in the blocked group, due to catastrophic forgetting of the first task (*bias:* $T(98) = 6.69, p < 0.0001$, *lapse:* $T(98) = 3.20, p = 0.0028$, *slope:* $T(98) = 0.02, p = 0.981$, *offset:* $T(98) = 0.02, p = 0.9826$, **Fig. 5.7B**). In contrast, fits to the sluggish Hebbian Gating network were similar to those observed in human data (*bias:* $T(98) = -3.76, p = 0.0006$, *lapse:* $T(98) = 0.06, p = 0.9555$, *slope:* $T(98) = 11.06, p < 0.0001$, *offset:* $T(98) = 1.32, p = 0.1964$, **Fig. 5.7C**). Taken together, these findings demonstrate how two adjustments to the training procedure, the introduction of sluggish task signals and a Hebbian learning step that is alternated with SGD updates, are sufficient to protect against catastrophic forgetting and model the cost of interleaved training observed in human participants.

5.2.6 Sluggish task estimates bias internal representations under interleaved learning

Why did the "sluggish" task signal lead to intrusions from irrelevant dimensions? In the original paper, we hypothesised that humans benefit from blocked training as it aids the formation of "factorised" representations, while interleaved learning might induce shared representations (Flesch et al., 2018). In subsequent neuroimaging work described in the previous chapter, we found evidence for such factorised and orthogonal representations in fronto-parietal areas of the human brain after blocked training. However, it is less clear how interleaved training might shape internal representations. We hypothesised that while blocked training should lead to orthogonal representations, interleaved training might induce representations that preferentially encode congruent stimuli, i.e., those that required the same response across tasks and lie on the main diagonal of the two-dimensional stimulus space.

To test this, we regressed RDMs from the hidden layer of the model trained with large values for the sluggishness parameter against a set of candidate RDMs encoding grid-like, "orthogonal", or "diagonal" representations. The grid model served as control and assumed that both feature dimensions were encoded in both tasks, forming a task-agnostic representation. In contrast, the orthogonal model represented the case where, starting from this grid model, task-irrelevant feature dimensions were filtered out, leaving a task-specific representation that encodes the relevant dimension in each context, with the two representations being orthogonal to each other. Lastly, in the diagonal model, representations of the stimuli were projected onto the main diagonal of the two-dimensional stimulus space which corresponded to stimuli that required the same response across tasks (methods). Indeed, while the orthogonal model explained the patterns best under blocked training (*grid vs orthogonal:* $T(98) = -347.62, p < 0.0001$*; orthogonal vs diagonal:* $T(98) = 178.67, p < 0.0001$), the diagonal model, which represented congruent stimuli, was the best predictor of hidden layer activity under interleaved training (*grid vs diagonal:* $T(98) = -109.23, p < 0.0001$*; orthogonal vs diagonal:* $T(98) = -73.89, p < 0.0001$, **Fig. 5.8A**). How were these representations formed? Assessing the task-selectivity of individual units in the hidden layer revealed

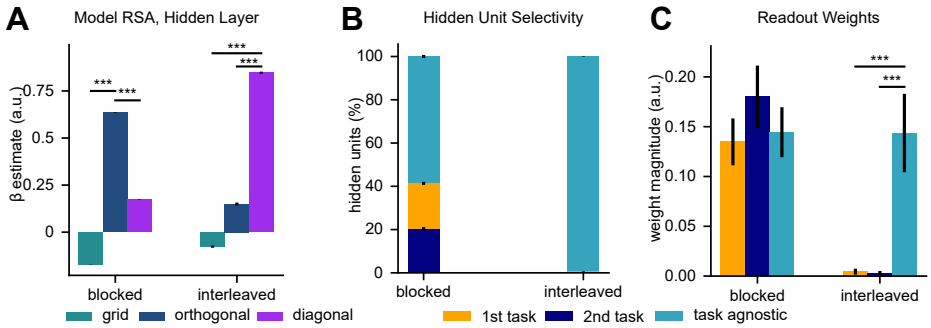

Figure 5.8: Sluggish task estimates under interleaved training bias internal representations. (**A**) Coefficients obtained by regressing hidden layer RDMs against three model RDMs, encoding grid, orthogonal or diagonal representations. The orthogonal model fits best under blocked training, whereas the diagonal model fits best under interleaved learning, suggesting that the latter aligns the representations with stimuli encountered in congruent trials. (**B**) Proportion of task selective units under both curricula. Under blocked learning, more units are task-selective than under interleaved learning. (**C**) Magnitude of readout weights, showing that under interleaved learning, the network relies more strongly on task-agnostic units (which encode congruent trials).

that while a sizeable fraction of units was selective to the relevant dimensions of each task under blocked learning (41.3%), most hidden units of the network trained on interleaved data were task-agnostic (99.4 %, **Fig. 5.8B**). Lastly, in the model trained on an interleaved curriculum, readout weights from those task-agnostic units were significantly larger than those reading out from the task-selective weights (*task agnostic vs 1st task:* $T(98) = 5.52, p < 0.0001$*; task agnostic vs 2nd task:* $T(98) = 5.58, p < 0.0001$, **Fig. 5.8C**). Together, these analyses suggest that due to the hypothesised sluggishness, interleaved training might not only alter the readout, but also the geometry of task-representations, providing avenues for further empirical research.

5.3 Discussion

Previous work has shown that humans perform worse after blocked compared to interleaved training on multiple categorisation tasks with high-dimensional stimuli (Carvalho & Goldstone, 2014; Flesch et al., 2018). In contrast, to converge, deep neural networks require training data to be randomly interleaved, as they suffer from catastrophic forgetting under blocked curricula. This limits both their performance and their viability as a model of human learning (Flesch et al., 2018; Parisi et al., 2019; Hadsell

et al., 2020).

Here, we propose a neural network model of human continual learning which captures this benefit of blocked over interleaved training and recreates several observations made in human participants at the behavioural and neural level. First, we demonstrated how a ”sluggish” task signal introduces biases in the acquired task representations which leads to worse performance under interleaved training. We note earlier reports that have previously proposed similar approaches to account for the cost of interleaving (Flesch et al., 2021; Russin et al., 2022). Secondly, we showed how gating, an inherent property of prefrontal cortex function, could not only be used to control switching between already learned tasks, but might indeed play an active role in the acquisition of novel tasks without forgetting. We propose that by augmenting standard supervised training with a Hebbian update, this gating scheme can be learned from scratch. Building directly on previous work on representation learning in humans and neural networks, we illustrated how these two properties shape neural representations, and how the emerging representational geometry can influence behaviour. Lastly, we validated our model by fitting it to previously published human behavioural data, allowing us to recreate the performance difference between blocked and interleaved training. Decomposition of these differences into different sources of error revealed that in both human participants and our model, differences were predominantly driven by a misestimation of the category boundary under interleaved training.

The idea that sluggish neurons could model costs associated with task switching is not new. In early models of cognitive control, it was assumed that the PFC has a bias to maintain task information over time (O’Reilly & Frank, 2006). This ”active maintenance” of task information would lead to intrusions between competing objectives and could explain why humans usually perform worse immediately after a switch to a different task (Herd et al., 2014). Here, we extended this idea and investigated how switch costs shape credit assignment during learning, demonstrating that interleaved training impairs the ability to link relevant perceptual information to the correct contextual cue.

A key component of our model is non-linear gating of internal representations. Early connectionist models have demonstrated how gating could be utilised by PFC

to minimise interference during multi-tasking (Cohen et al., 1990; Miller & Cohen, 2001), and follow-up work suggested that basal ganglia could control the gating of PFC representations (O'Reilly & Frank, 2006). However, with few exceptions such as (Iyer et al., 2022; Rougier et al., 2005) and (Flesch et al., 2022), the gating was usually hand-crafted by the experimenters, and it remained unclear how these control processes might emerge in the first place (Verbeke & Verguts, 2022). Similarly, a handful of studies have drawn the link to continual learning and investigated how gating could prevent catastrophic forgetting, but once again, the process was usually implemented by hand (Masse et al., 2018; Russin et al., 2022; Tsuda et al., 2020). We demonstrate that a simple biologically inspired intervention (Hebbian learning) is sufficient to implement this gating strategy.

At a representational level, the gating effectively orthogonalises hidden layer representations by enforcing an axis-aligned coding scheme. Interestingly, a recent series of papers has provided converging evidence that the brain might use orthogonal representations to minimise interference between tasks (Flesch et al., 2022; Libby & Buschman, 2021; Panichello & Buschman, 2021; Xie et al., 2022) and some of the more successful recent engineering solutions to Catastrophic Forgetting employ orthogonalization of gradient updates of internal representations (Chaudhry et al., 2020; Farajtabar et al., 2019; Zeng et al., 2019). Here, we propose a biologically inspired model of how these orthogonal representations could be learned.

Our model appears to be strongly related to a recently published conference submission in which the authors demonstrated that carrying over task-signals from previous trials leads to lower performance on interleaved curricula (Russin et al., 2022). Like in our work, the authors propose an implementation of "sluggishness" that is inspired by models of switch costs in PFC and suggest a simple gating mechanism to prevent forgetting. However, while the authors implemented this gating scheme manually, we propose a Hebbian training step that can learn this scheme from scratch. Leaving differences in implementational details aside, both studies provide converging evidence that theories on the role of PFC for cognitive control can be readily extended to the problem of continual learning.

There are several clear avenues for future research. We introduced the notion of sluggishness to account for performance costs observed in human participants under interleaved training. Similar to other recent accounts, we assumed this sluggishness to be an inherent property of prefrontal function (Russin et al., 2022). Future work could investigate the normative basis of such a coding scheme. For example, under blocked training, the active maintenance of task signals might protect against noise in the task signal. Under this account, sluggishness would ensure ongoing task performance under blocked curricula, even if the task signal could not be read out or was mislabelled on a subset of trials. An even stronger claim, building on previous work on sequential effects in human decision making (Yu & Cohen, 2008), would be that sluggishness might adapt to the volatility of the environment. Future work could investigate if the window over which contextual information is averaged depends on the amount of time spent in a single context, or the extent to which task switches are predictable from recent trial history. It should also be noted that blocked training is not always advantageous, as there seem to be several cases in which humans benefit indeed from interleaved curricula (Rohrer et al., 2015; Samani & Pan, 2021). In our simulations, we observed that at the level of hidden units, sluggish task signals promote the formation of shared representations that don't arbitrate among tasks. A prediction that arises from these simulations is that sluggish units might help the learner to find similarities among tasks that are encountered in close temporal proximity. Consequently, it is likely that whether sluggishness introduces a cost or benefit for learning depends on the similarity between tasks and their transfer demands, and hence the need for shared or separated representations (Musslick et al., 2020; Musslick & Cohen, 2021).

Another possible line of enquiry is lifelong learning. We focused on a simple and tractable context-dependent decision-making problem with only two tasks. Our use of Hebbian updates was motivated by the connection between Oja's rule and PCA, as it picks up the direction of highest variance, which happened to be spanned by the zero-centred task signal in our dataset. Future work could investigate how this approach extends to additional tasks, both at the human behavioural level and in artificial neural networks. We note that our Hebbian procedure is in essence achieving a temporal

clustering of contextual information, with the active cluster gating on a set of units and inhibiting the rest. This scheme in principle might work in richer settings with additional tasks. Real lifelong/continual learning, however, is likely to involve more than a single Hebbian learning mechanism applied to prefrontal gating signals. Future work could investigate how these gating processes interact with memory consolidation and replay of previous experiences during sleep.

To conclude, we introduced two algorithmic motifs to augment vanilla neural networks trained with stochastic gradient descent, "sluggish" task signals and a Hebbian update step, which together are sufficient to model the benefit of blocked over interleaved training previously observed in humans. Furthermore, investigation of the learned representations suggests that blocked training might promote the formation of orthogonal representations, like those observed in biological brains, while interleaved training leads to shared representations that optimise for congruent trials. Taken together, we provide a biologically inspired model of human continual learning, grounded in previous work on representation learning and the function of prefrontal cortex.

5.4 Methods

5.4.1 Software

All simulations were implemented in Python 3.9 with the PyTorch 1.71 package. Hyperparameter tuning was carried out with the RayTune 1.10 package. Parallelisation was implemented with the joblib 1.0.1 package. Stimuli were generated in Python with the NumPy 1.19.2 and SciPy 1.60 packages. All statistical analyses were carried out in Python with the Pandas 1.2.3, NumPy 1.19, SciPy 1.60, Statsmodels 0.13 and Scikit-Learn 0.24.1 packages. Figures were generated with Matplotlib 3.3.2.

5.4.2 Experimental Design

Stimulus design: Stimuli were grayscale images of two-dimensional Gaussian functions with isotropic covariance. We varied the mean of these Gaussian "blobs" in five discrete steps along the x- and y-coordinate, creating a 5x5 grid of possible stimulus locations inside these image patches. The Gaussian blobs were partially overlapping.

This gave the network some information about the two-dimensional structure of the stimulus space, which would not have been the case with a conventional one-hot encoding of stimuli.

Task design: We trained feedforward neural networks on a context-dependent decision-making problem, where only a single dimension of the Gaussian blobs (either the x-or y-location) was relevant for each task/context. Each task was defined by a category boundary that divided this space either along the horizontal (first task) or vertical axis (second task). In each task, the network had to learn to "accept" stimuli from one category and "reject" stimuli from the other category.

Neural network architecture: For all simulations, we used a feed-forward neural network with 25 input units (for the flattened and downscaled grayscale images) and two additional task units, a hidden layer with 100 Rectified Linear Unit (ReLU) nonlinearities and a sigmoidal output unit. Weights from the input to the hidden layer were initialised with draws from a zero-mean Gaussian distribution with variance $\sigma^2 = 0.01$. Readout weights were initialised with draws from a zero-mean Gaussian with variance $\sigma^2 = \frac{1}{\sqrt{n\ hidden}}$. All biases were initialised to zero.

5.4.3 Training Procedures

All networks were trained on 10000 trials, 5000 per task. In the interleaved curriculum, trials from both tasks were randomly shuffled. In the blocked curriculum, the networks were first trained on all 500 trials from one task, and then on all trials from the other task. Following our previous publication (Flesch et al., 2018), we used a custom loss function which was -1 times the reward associated with "accepting" a Gaussian blob. This was implemented by multiplying the output of the network function (which was in the range 0 to 1 due to the sigmoid) with -R:

$$J(\theta) = -f(x, \theta) R \tag{5.1}$$

Rewards ranged from -2 to 2 in steps of 1, hence covering all 5 levels of the feature value along the relevant dimension. Hence, the network was encouraged to "accept" rewarding and "reject" non-rewarding stimuli. At the end of the training phase, we

evaluated the network's performance on 50 test trials spanning all combinations of task (2), x-position (5) and y-position (5) of the stimuli. For each simulation, we collected 50 independent training runs with randomly initialised neural networks.

Baseline model: The baseline network was trained with vanilla Gradient Descent, applied via Backpropagation to all network weights after each trial:

$$W^{t+1} \leftarrow W^t - \varepsilon \nabla^{(W^t)} J(x_t, W) \tag{5.2}$$

with a learning rate of $\varepsilon = 0.2$ for the interleaved and $\varepsilon = 0.03$ for the blocked curriculum.

Sluggishness: We modelled the "sluggishness" property of the task signals with an Exponentially Moving Average (EMA), which was applied to the task units on each trial. The EMA has the following recursive definition:

$$x_t^{EMA} = \begin{cases} x_1 & for\ t = 1 \\ (1-\alpha)x_t + \alpha x_{t-1}^{EMA} & for\ t > 1 \end{cases} \tag{5.3}$$

where the hyperparameter α controls the extent to which information from previous trials is carried over to the current trial. To investigate the impact of the sluggishness on task performance, we trained the baseline model (see above) on an interleaved curriculum for a linearly spaced range of 20 α values ranging from 0 to 0.95 and a fixed learning rate of $\varepsilon = 0.2$. We collected 50 independent training runs with randomly initialised networks for each of these values.

Continual learning with manual gating: To investigate the impact of non-linear gating on continual task performance, we manually set the weights connecting the task units with each hidden unit to values with opposing signs. More specifically, all "odd" hidden units received a negative bias in the first task and positive bias in the second, whereas all "even" hidden units received the opposite:

$$w_i^h = \begin{cases} [1,-1] & for\ i \in \{1,3,5\ldots,n-1\} \\ [-1,1] & for\ i \in \{2,4,6,\ldots,n\} \end{cases} \tag{5.4}$$

We trained the remaining weights of the network with vanilla SGD, just as described for the baseline model above. The learning rate was set to $\varepsilon = 0.01$. The network was trained on a blocked curriculum, and we collected 50 independent training runs.

Continual learning with Hebbian updates and SGD: To protect against interference under blocked training, we devised a novel training procedure which consisted of alternating the standard SGD update and a Hebbian learning step. The Hebbian update enabled the network to strengthen associations between the task units and hidden units that carried task-relevant information, while suppressing the output of units with task-irrelevant information. In the following, we motivate this solution from first principles. Hebbian learning strengthens connections between units that are co-activated. Given inputs x and linear hidden units y connected to the inputs via weight matrix W as follows:

$$y_j(x) = \sum_{i=1}^{n} w_i x_i = w_j^T x \tag{5.5}$$

Hebbian learning performs weight updates proportional to the co-activation of x_i and y_j:

$$\Delta w_{ij} = \eta x_i y_j = \eta x_i \sum_{i=1}^{n} w_i x_i \tag{5.6}$$

or for the entire vector of weights from inputs to a single hidden unit:

$$\Delta w_j = \eta x y_j = \eta x \sum_{i=1}^{n} w_i x_i \tag{5.7}$$

The weight updates for standard Hebbian learning are unbounded, which means that weights continue to grow as training progresses. A conventional solution to this problem is to introduce weight decay, leading to the well-known Oja's rule (Oja, 1982):

$$\Delta w_j = \eta y_j \left(x - w_j y_j\right) \tag{5.8}$$

Oja's rule converges to the first principal component of the dataset, such that w encodes the first eigenvector and y the first eigenvalue of the input covariance matrix (Oja, 1982). This can be seen by slightly rearranging the terms. First, in the classical formulation of Hebbian learning, we set the learning rate η to 1 and introduce an

average over multiple trials:

$$\Delta w_i = \langle y x_i \rangle \tag{5.9}$$

$$\Delta w_i = \langle \sum_j^n w_j x_j x_i \rangle = \sum_j^n w_j \langle x_j x_i \rangle \tag{5.10}$$

$$\Delta w = \left\langle x x^T \right\rangle w = C w \tag{5.11}$$

With this formulation, the growth of weights w depends solely on the input-input correlation matrix C. Now recall that the update equation for Oja's rule is given by

$$\Delta w = y\,(x - wy) = yx - y^2 w \tag{5.12}$$

Introducing the average over multiple examples yields

$$\Delta w = \langle yx \rangle - \left\langle y^2 \right\rangle w \tag{5.13}$$

The equilibrium for this equation is reached when the first term on the right is equal to the second, or in other words, when:

$$\langle yx \rangle = \langle y^2 \rangle w \tag{5.14}$$

$$Cw = \langle y^2 \rangle w \tag{5.15}$$

From the definition of eigenvalues, it follows that w is an eigenvector of C and $\left\langle y^2 \right\rangle = \sigma^2$ its corresponding eigenvalue. Further, the dynamics grow fastest in the direction of the eigenvector with maximal eigenvalue, such that w will converge to the largest principal component of the input data. Applied to the blobs task, this means that weights from the task units to some of the hidden units are positive for one task and negative for the other, while the opposite is true for other units. Together with the supervised learning step, this should allow the network to strengthen positive weights between the active task units and task-relevant hidden units, and negative weights between this task unit and task-irrelevant hidden units. Once the network is exposed to a

new task, the opposite mapping should be learned for connections between the second task unit and the hidden layer. We implemented this procedure as follows. For each trial and corresponding input sample x_t, we first applied the standard SGD update via backpropagation to all network parameters:

$$W^{t+1} \leftarrow W^t - \varepsilon \nabla^{(W^t)} J\left(x_t, W\right) \tag{5.16}$$

This was then followed by a Hebbian update to the weights from the task units to the hidden layer, where y corresponds to the hidden layer activation of the j-th hidden unit prior to the non-linearity and each w_j corresponds to a vector of weights from all task-units to the $j-th$ hidden unit:

$$w_j^{t+1} \leftarrow w_j^t + \eta y_j \left(x_i - w_j^t y_j\right) \tag{5.17}$$

We trained the network on a blocked curriculum as described above, with a learning rate of $\varepsilon = 0.03$ for the SGD and $\eta = 0.05$ for Hebbian updates with Oja's rule. We collected 50 training runs with independent random initialisations of all network parameters. One might object that mean-centring the task-signal introduces knowledge about the second task during training on the first, as the one-hot inputs $[1,0]$ were converted to $[0.5,-0.5]$. To overcome this, we used a one-hot signal for the first task and introduced a mean-centred signal for the second task during training. Semantically, this would correspond to first learning how to perform the first task, and then how to do the second task while suppressing information learned about the first.

Modelling human continual learning: To model human continual learning, we combined the sluggishness and Hebbian update procedure outlined above as follows: On each trial, The task signal received by the network was mixed with the signal carried over from previous trials:

$$x_t^{EMA} = \begin{cases} x_1 & for\ t = 1 \\ (1-\alpha)\, x_t + \alpha x_{t-1}^{EMA} & for\ t > 1 \end{cases} \tag{5.18}$$

Next, we performed a forward pass through the network and calculated the loss

as -R:

$$J(W) = -\sigma(x,W)R \tag{5.19}$$

This was then used to perform an SGD update of the network parameters, with a learning rate of $\varepsilon = 0.03$ for the blocked curriculum and $\varepsilon = 0.03$ for the interleaved curriculum:

$$W^{t+1} \leftarrow W^t - \varepsilon \nabla^{(W^t)} J(x_t, W) \tag{5.20}$$

Lastly, the task weights were updated with Oja's rule, with a learning rate of $\eta = 0.05$ for the blocked and $\eta = 0.05$ for the interleaved curriculum:

$$w_j^{t+1} \leftarrow w_j^t + \eta y_j \left(x_i - w_j^t y_j\right) \tag{5.21}$$

In contrast to the neural network model, Human participants never performed at ceiling on test trials with novel stimuli, not even after extensive training on the tasks. To model this residual cost, we introduced decision noise at test by passing the network's logits through a sigmoid with temperature parameter T that controlled its sensitivity to changes in the input:

$$\sigma(x) = \frac{1}{1 - exp\left(-\frac{x}{T}\right)} \tag{5.22}$$

At test, we sampled 10000 choices per input from the trained model by comparing its output to a random uniform variable $X\ U(0,1)$:

$$y = \begin{cases} 1 & for \sigma(x,W) \geq X \\ 0 & for \sigma(x,W) < X \end{cases} \tag{5.23}$$

To fit this model to human choices, we performed a grid search over a range of values for the alpha and T parameters that controlled the amount of sluggishness and decision noise respectively and chose those values that produced outputs which closely resembled the choices made by human participants.

5.4.4 Quantification and statistical analyses

Test accuracy: To compute accuracy during training and test, we evaluated whether the network accepted the rewarding and rejected the non-rewarding trials. Excluding

the boundary trials for which the decisions were arbitrary, accuracy was calculated as follows:

$$p\,(correct) = \frac{1}{n}\sum_{i}\mathbf{1}_{f(x_i,W)>0.5==R_i>0} \tag{5.24}$$

Choice matrices: To visualise the choices made by the network, we averaged outputs across trials for each of the 50 unique types of test trials (5 x-positions, 5 y-positions, 2 tasks) and rearranged these outputs into two 5x5 matrices where each entry corresponds to the fraction of "accept" responses for this type of stimulus.

Task selectivity: We performed a regression-based analysis to determine task-selectivity of individual neurons. We regressed their activity against four predictors, coding for the value of relevant and irrelevant feature dimensions of each trial, separately for each task:

$$y_{unit} = \beta_0 + \beta_1\,\text{relDim}_{1st\ task} + \beta_2\,\text{irrelDim}_{1st\ task} + \beta_3\,\text{relDim}_{2nd\ task} + \beta_4\text{irrelDim}_{2nd\ task} \tag{5.25}$$

Following procedures explained in detail in (Flesch et al., 2022), we defined a unit as being task-selective if its output scaled with the feature value along the relevant – but not irrelevant - dimension of one task, and was zero for the other task. This definition results directly from the rectifying property of ReLUs, which are linear for positive inputs and return zero for negative inputs. It only counts those units as task-selective that have receptive fields aligned with task-relevant information and doesn't consider units that happen to be active in one task, but not the other.

Hidden layer Representational Similarity Analysis (RSA): We performed representational similarity analysis (RSA) to investigate the geometry of hidden layer activity patterns of the trained neural networks. First, we collected activity patterns for all 50 conditions (5 x-positions, 5 y-positions, 2 tasks), yielding a 50-x-n_hidden matrix of activity patterns for each individual training run. Next, we created 50x50 representational dissimilarity matrices (RDMs) by computing the pairwise Euclidean distance between all 50 patterns. For visualisation purposes, we then averaged these RDMs across training runs (separately for the blocked and interleaved curriculum) and projected them down into 3 dimensions using classical Multi-Dimensional Scaling (MDS). As MDS

is rotation-invariant, we manually rotated the resulting projection so that axes of the projection were aligned with the figure axes, which made it easier to compare the geometry across conditions (and models). To get quantitative insights into the geometry of these patterns, we regressed these RDMs against a set of model RDMs that encoded (a) grid-like, (b) orthogonal or (c) diagonal patterns at the level of individual runs. As these models were identical to the ones described in the previous two chapters, the interested reader is referred to their corresponding methods sections. To estimate the extent to which each of these models explained the geometry of representations in the hidden layers of our neural networks, we performed a multiple linear regression at the level of individual runs, in which we regressed the hidden layer RDM against the set of model RDMs, after z-scoring and vectorising the lower-triangular form of each RDM:

$$\text{RDM}_{hidden} = \beta_0 + \beta_1 \text{RDM}_{grid} + \beta_2 \text{RDM}_{orth} + \beta_3 \text{RDM}_{diag} \tag{5.26}$$

For statistical inference at the group-level, we performed t-tests against zero on each set of regression coefficients.

Comparison with human behavioural data. We followed procedures described in (Flesch et al., 2018) for our re-analysis of the behavioural data. In the original study, there were four groups that differed in the amount of "blockiness" during training, ranging from a fully blocked curriculum where participants were trained on one task and then the other, to a fully interleaved curriculum in which trials were randomly interspersed. In our re-analysis, we focus on the two extremes, called the "blocked 200" group and "interleaved" group in the original publication. Detailed descriptions of the calculation of the sigmoidal fits, model-based RSA and fits of the psychophysical model are provided in the methods section of chapter 4, where we applied identical procedures.

Chapter 6

Blocked versus interleaved training for cross-domain transfer

Abstract

Throughout our lifetime, we continually develop a sophisticated understanding of the world around us. Not only are we able to learn multiple tasks in succession without forgetting, we can also generalise knowledge between related contexts. Here, we investigated whether blocked, in contrast to interleaved training, conferred a benefit for cross-domain transfer. We first tested whether a previously reported benefit of blocked over interleaved training for context-dependent decisions generalises to abstract stimulus features. For this, human participants had to learn the size and speed dimensions of naturalistic images of either vehicles and animals. Next, we assessed transfer performance in a new cohort of participants that was trained either on animals or vehicles and tested on the held-out domain. Across all studies, we observed only very limited evidence for a benefit of blocking over interleaving. Analyses of the choice patterns and participant reports indicates that our participants performed either rote-learning of stimulus-response contingencies, or relied on dimensions other than size and speed to inform their choices.

6.1 Introduction

Humans have the remarkable ability to learn multiple tasks throughout their lifetime. In the previous chapters, we explored the coding schemes that might enable humans

to learn without catastrophic forms of forgetting and found evidence for orthogonal task representations which minimise interference between tasks. Another hallmark of human cognition that sets us apart from state-of-the-art AI approaches is the ability to re-use information in related contexts (Ferguson, 1956; Lake et al., 2016). For example, rather than representing the speed of animals and the speed of vehicles as completely separate concepts, we understand that speed is a domain-general, abstract concept that can be applied to animals and vehicles alike (Baram et al., 2021; Behrens et al., 2018; Sheahan et al., 2021; Summerfield et al., 2020). The conditions under which we can learn mappings onto these domain-general representations, however, are only poorly understood.

In a previous study, we demonstrated that blocked, in contrast to interleaved training on context-dependent decision tasks allows humans to learn factorised representations with accurate estimates of the context-specific category boundaries (Flesch et al., 2018). Under interleaved training on the other hand, participants tended to learn a single boundary which optimised for performance across contexts. How these findings generalise to semantic categories, however, remained an open question. Learning rules with naturalistic stimuli and semantic categories are particularly challenging, as these stimuli can theoretically vary along a multitude of dimensions, both perceptually and in the semantic domain, and memory recall is required for successful categorisation. Earlier reports indicate that humans might indeed benefit from blocked training under these circumstances, as it helps with the detection of task-relevant structure in tasks with large within-category variance, while interleaved training might be beneficial when there is only little variation within categories (Carvalho & Goldstone, 2014; Kurtz & Hovland, 1956).

Here, we report two experiments in which we tested whether this benefit of blocking over interleaving extends to abstract, semantic categories, and whether representations learned in such a regime would generalise to a second domain with shared task structure. In the first experiment, we trained participants on the speed and size of either animals or vehicles in a well-established context-dependent decision paradigm (Flesch et al., 2018; Mante et al., 2013) and assessed their performance on a subse-

quent interleaved test phase with stimuli from the same domain. Participants were not aware that stimuli varied systematically in size and speed and had to identify these as task-relevant dimensions from trial-wise feedback alone. In the second experiment, we tested whether training on either blocked or interleaved trials from one domain (e.g. animals) would facilitate transfer to another domain (e.g. vehicles) with the same task-relevant dimensions.

Across the two experiments, earlier results replicated only partially. While results from experiment 1 suggested that participants trained on blocked trials could indeed learn more accurate estimates of the category boundaries, this effect did not generally replicate in the second experiment. Analyses of the choice patterns and written reports provided by our participants after completion of the experiment, however, suggests that they relied on alternate strategies, such as rote-learning or stimulus dimensions other than size and speed, to perform the tasks.

6.2 Results

We report results from two experiments. In experiment 1, we assessed whether the previously reported benefit of blocking over interleaving for context-dependent decisions on perceptually varying stimuli could be replicated with semantic stimulus spaces. In experiment 2, we tested whether blocked, in contrast to interleaved learning would promote cross-domain transfer in these semantic spaces. To address these questions, we curated stimulus sets of naturalistic images depicting animals or vehicles that varied in their size and speed (how fast they can run/move) in five discrete steps, spanning a two-dimensional space of 25 different feature combinations per domain (animals/vehicles) (**Fig. 6.1A-B**). Prior to running the main experiments, we normed these stimuli with a separate rating task in which we asked different groups of participants to rate the speed/size of animals/vehicles on a five-point Likert scale and assessed the agreement between their ratings and our ground-truth labels (**Appendix C, Supplementary Methods, Fig. C.1,C.2,C.3**).

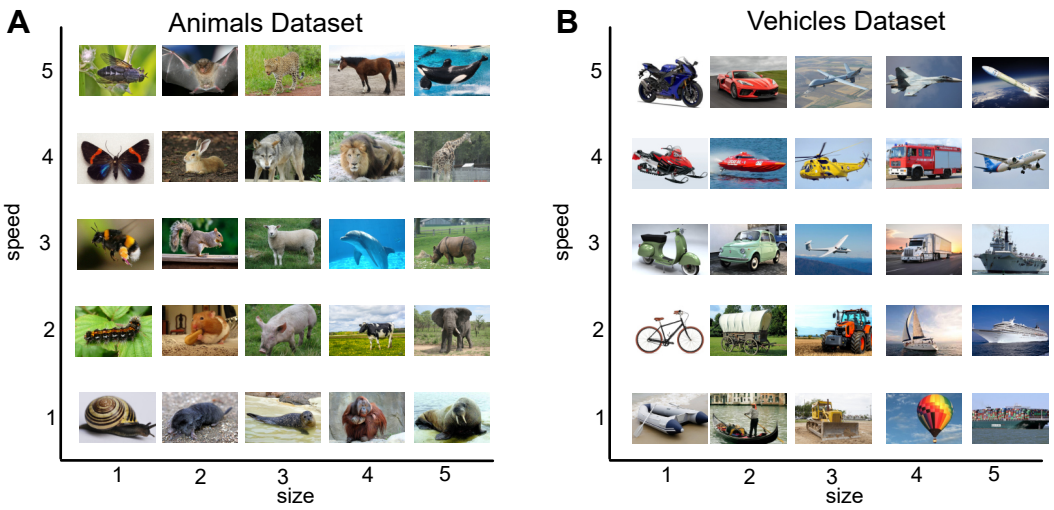

Figure 6.1: Stimulus Spaces. (**A**). Animals dataset. Stimuli were images of animals that varied in speed and size along five discrete steps. (**B**) Vehicles dataset. Same as (A), but with images of vehicles.

6.2.1 The blocking benefit replicates only partially in semantic spaces

In experiment 1, we trained separate groups of participants either on the animals dataset (experiment 1a), or the vehicle dataset (experiment 1b). In both cases, participants were asked to learn what kind of animals/vehicles would be preferred by customers in two different shops, the orange and blue shop respectively, and accept only those that would yield a reward (indicating the preference) (Fig. 6.2B). Unbeknownst to the participants, these preferences dependent on one of two semantic feature dimensions per shop, the speed of animals/vehicles in the blue shop and the size of animals/vehicles in the orange shop (**Fig. 6.2A,C**). Hence, participants had to identify task-relevant semantic dimensions and learn their mapping onto penalties/rewards for accepting/rejecting stimuli in each shop. Crucially, they were either subjected to a blocked training curriculum with one block per task, or a fully interleaved curriculum in which context switches occurred randomly. Both participant groups were evaluated on an interleaved test phase without feedback (**Fig. 6.2D**).

While our participants were able to learn the tasks, reaching 75% training accuracy both on the animals and the vehicles dataset (**Fig. 6.3A**), we observed no significant difference in test accuracy between blocked and interleaved training (*An-*

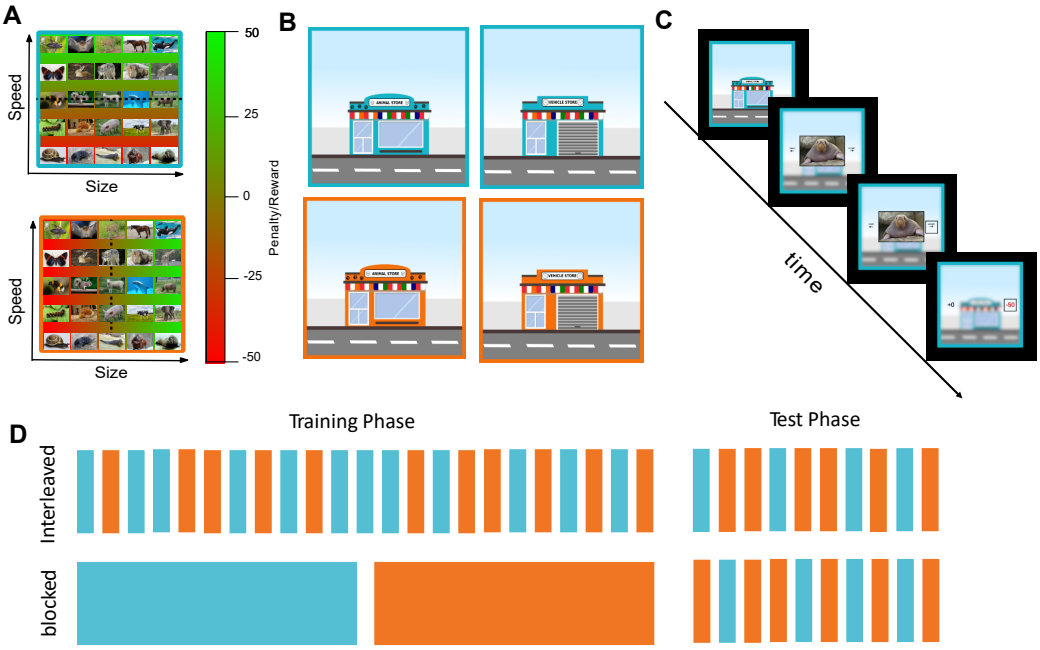

Figure 6.2: Task design, experiment 1. (**A**) Mapping from feature dimensions (size/speed) onto rewards for accepting an animal in each context (orange/blue). Only one dimension was relevant per context. Image illustrates mapping for experiment 1a. The same mapping was used for experiment 1b (vehicle dataset). (**B**) Images of shops that served as contextual cues on each trial. Left column: Shops used in experiment 1a (animals). Right column: shops used in experiment 1b (vehicles). (**C**) Example trial sequence. Participants saw an image of a shop, followed by an image of an animal (experiment 1a) or vehicle (experiment 1b) which they could accept or reject. On training trials (shown), feedback for the chosen and unchosen option was displayed. No feedback was provided on test trials (see methods for details). (**D**) Training curricula. In both experiments, participants were either trained on a "blocked" or an "interleavead" curriculum. All participants were evaluated on interleaved trials without feedback in a subsequent test phase.

imals: $T(177) = 1.271 p = 0.205$*; Vehicles:* $T(183) = -0.189 p = 0.850$, **Fig. 6.3A**). Similarly, in terms of sensitivity of choices to variation along the relevant dimensions, there was no difference between groups, neither in the slope nor in the offset of sigmoids fit to participant's choice patterns (*Animals: slope (relevant)* $T(177) = 1.85, p = 0.066$*, offset (relevant)* $T(177) = 0.442, p = 0.659$*; Vehicles: Slope (relevant)* $T(183) = 1.016, p = 0.311$*, offset (relevant)* $T(183) = -0.72, p = 0.473$, **Fig. 6.3B**). Intrusions from the irrelevant dimension, however, were larger under interleaved compared to blocked training in the animals dataset, suggesting that error patterns might have at least in part differed between groups (*Animals: slope (irrele-*

vant) $T(177) = -2.487, p = 0.014$*, offset (irrelevant)* $T(177) = 1.48, p = 0.141$*; Vehicles: Slope (irrelevant)* $T(183) = -0.008, p = 0.993$*, offset (irrelevant)* $T(183) = -0.891, p = 0.374$, **Fig. 6.3B**).

Previously, we had reported that interleaved training leads to higher errors in the estimation of the category boundary, with a tendency towards a "linear" solution, in which the same category boundary is applied in both tasks. We contrasted this with a "factorised" model, in which separate and highly accurate category boundaries are learned (Flesch et al., 2018). Here, we replicate this finding, with better fits for the linear model under interleaved training in both dataset (*Animals, linear model, interleaved > blocked:* $T(177) = 2 - 3.398, p = 0.001$*; Vehicles, linear model, interleaved>blocked:* $T(183) = -2.785, p = 0.006$, **Fig. 6.3C-D**). As in our previous report, at least for the animals dataset, the factorised model fit better to the blocked compared to the interleaved group (*Animals, factorised model, blocked>interleaved* $T(177) = 2.898, p = 0.004$*; Vehicles, factorised model, blocked>interleaved* $T(181) = -1.128, p = 0.261$, **Fig. 6.3C-D**).

If choice patterns really differed between groups, why could we not detect a difference in overall accuracy at test? Previously, we observed that for tasks with rules that are not cleanly mapped onto perceptually varying dimensions, blocking led to better estimates of the boundary, but this benefit was dampened by overall higher rates of unspecific errors ("lapses") (Flesch et al., 2018). To test whether this could explain the absence of overall accuracy differences, we fit a psychometric model to the choice patterns which we had developed for the previous study, with free parameters for the angle of the category boundaries in each task, as well as the slope, offset and lapse rate of a sigmoid non-linearity. While this analysis confirmed that boundary estimates differed between groups, with significantly lower errors in the animals dataset, but only a non-significant trend for vehicles (*Animals, bias in decision boundary, blocked < interleaved* $T(177) = -3.77, p = 0.001$*; vehicles, bias, blocked<interleaved* $T(183) = -1.596, p = 0.112$, **Fig. 6.4A**), lapse rates were similar under blocked and interleaved training (*Animals, lapse rate blocked>interleaved* $T(177) = -0.538, p = 0.594$*; vehicles, lapse rate blocked>interleaved* $T(183) = 0.483, p = 0.629$, **Fig. 6.4A**).

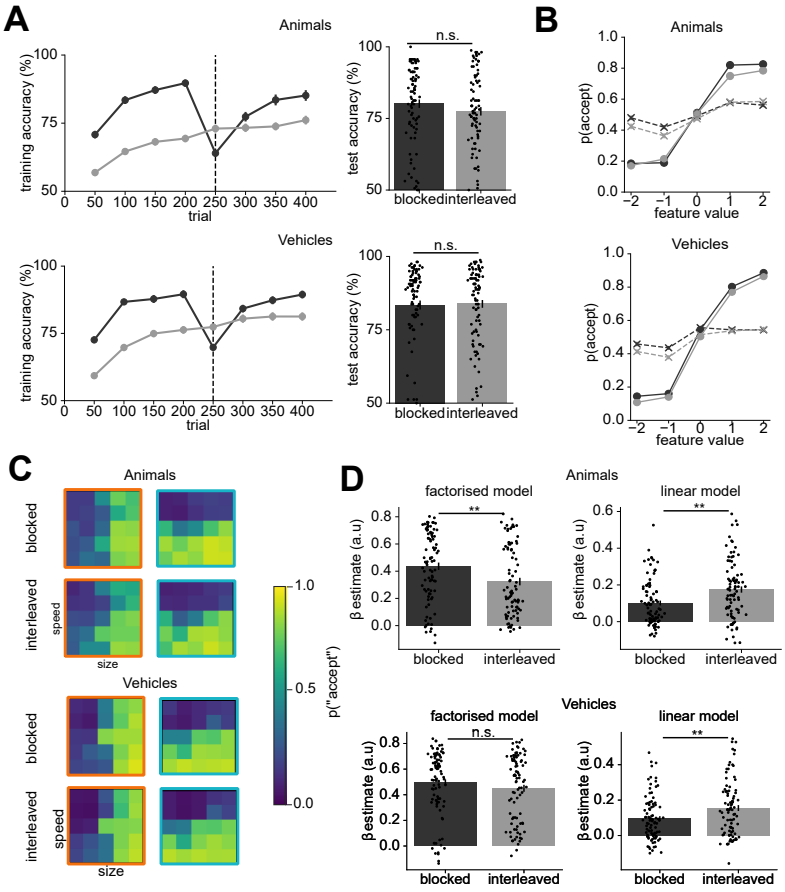

Figure 6.3: Behavioural results, experiment 1. (**A**) learning curves and test phase accuracy, shown for the animals (top) and vehicles (bottom) participant groups. Performance improved over the time course of the trial, but did not differ between training curricula. (**B**) Choice probabilities as functions of the feature value along the relevant (solid lines) and irrelevant (dashed lines) dimensions, shown separately for the animals (top) and vehicles (bottom) participant groups. Choices were more sensitive to the relevant than the irrelevant dimension, but intrusions from the irrelevant dimension were slightly larger under interleaved learning for the animals dataset. (**C**) Average choice patterns per shop. Across all tasks (orange/blue), curricula(blocked/interleaved) and domains (animals/vehicles), choice patterns were well aligned with the ground truth. Under interleaved training, there seemed to be a tendency toward a diagonal boundary. (**D**) Fits of linear and factorised models (methods) to the choice patterns shown in (C) on single subject level. The linear model explained the data better under interleaved compared to blocked training, suggesting that participants tended to a single category boundary. The factorised model fit better under blocked learning for the animals, but not the vehicles dataset.

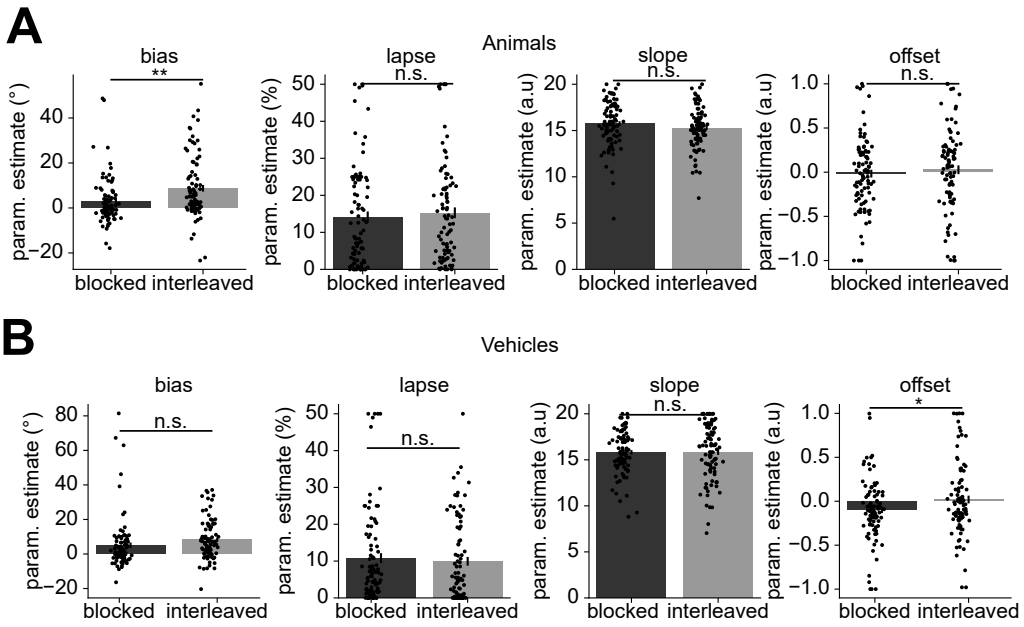

Figure 6.4: Psychophysical model fits, experiment 1. (**A-B**) Parameter estimates of psychophysical model, fit separately to participants from the blocked and interleaved groups of the animals (A) and vehicles (B) domains. Parameter estimates did not differ between curricula, with the exception of the angular bias (disparity between ground truth and estimated category boundary), which was smaller under blocked training on the animals dataset. All error bars denote SEM. Dots correspond to individual participants.

6.2.2 Topic modelling suggests that participants relied on alternative feature dimensions

In contrast to previously used stimuli that varied systematically along a small set of perceptually distinct dimensions, our animals and vehicles dataset might have features other than size and speed that were at least in part predictive of the reward in each shop. While participants in all groups improved their performance over the timecourse of the experiment, it remained unclear whether they had indeed based their decisions on size and speed or used alternative strategies. To gain insights into the rules our participants had identified, we asked them to report the strategies they had used at the end of the main experiment. As a first exploratory analysis, we looked at the overall frequency with which words appeared in the reports, separately for each combination of domain, curriculum, and task/shop. Indeed, both for the animals and vehicles dataset, words describing size, such as “large”, “big” or “small” were

very frequent in descriptions of the rules for the size tasks (*animals-blocked-size: 46, animals-interleaved-size: 43, vehicles-blocked-size: 23, vehicles-interleaved-size:43*, **Fig. 6.5A-B**). Similarly, words describing speed, such as "fast" or "slow" were frequently reported for the speed task, but much less often for the animals compared to the vehicles tasks (*animals-blocked-speed: <5, animals-interleaved-speed: <5, vehicles-blocked-speed:29, vehicles-interleaved-speeed: 30*, **Fig. 6.5A-B**). Overall, participants rather tended to give concrete examples for stimuli that were rewarding/salient (**Fig. 6.5A-B**).

How then, did our participants solve the task? We hypothesised that there might be dimensions other than size or speed which we weren't aware of when we designed the stimulus spaces, and that these could possibly explain the observed error patterns. To infer these feature dimensions from the written reports, we used topic modelling, a technique from natural language processing. In brief, topic modelling reveals clusters, so-called "topics" in a distribution of documents. In our case, these documents were individual reports and topics would group these reports by the described strategies. We plotted histograms for the number of participants whose reports were assigned to the top five latent topics, together with the top six words associated with each topic. While topics were broadly consistent across training regimes, this analysis revealed various alternative strategies that could have been used to solve the task. Overall, the results suggests that participants relied on a combination of various features, rather than a single feature per task. For animals, many participants indeed identified size as relevant dimension in the size task, but they also considered whether an animal was a mammal, or where it was predominantly found (land/air/sea). Interestingly, most participants distinguished between domestic and exotic/"zoo" animals (**Fig. 6.6A-B**). For the speed task, results were more heterogenous, with number of legs or whether it had wings being popular strategies. Interestingly, 27 participants from the blocked group reported that they relied on memorisation to solve the task, indicating that they learned individual stimulus-response mappings instead of discovering latent features (**Fig. 6.6A-B**). Results were similar for vehicles, where participants relied on the speed, but also on the perceived value of vehicles to inform their judgements in the speed task. In the size

task, they mainly distinguished between small and large vehicles, but also between those that carried cargo or passengers (**Fig. 6.6A-B**). Crucially, except for the size task in the vehicle domain, none of the most frequent topics in these reports mentioned the correct ground-truth feature (size/speed). While these analyses are highly exploratory, they indicate that our participants could have relied on a variety of semantic feature dimensions to solve the task. It should be noted, however, that these subject reports may still at best provide an indirect window into the true strategies used, as they rely on the ability of participants to memorise and verbalise the rules.

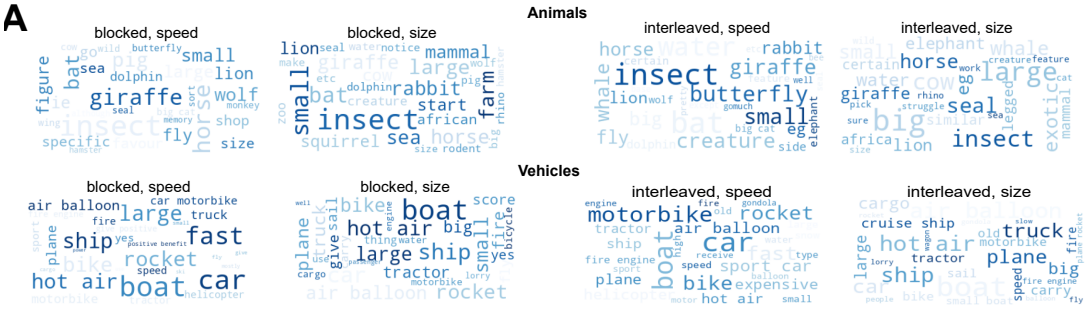

Figure 6.5: Word clouds, experiment 1. (**A**) Word clouds for animals task, showing frequency of the 30 most frequent words in the participant reports for each task, as indicated by their font-size. While words describing size ("large", "big", "small") occurred frequently in reports of the rules used for the size tasks, words describing speed ("fast", "slow") were much less frequent in the speed task. Overall, participants tended to give concrete examples for stimuli they chose to accept. (**B**) Same as (A) but for vehicles

6.2.3 No consistent difference between curricula in facilitation of cross-domain transfer

Next, we sought to test whether representations learned under blocked and interleaved training are domain-specific or domain-general. In theory, participants who had learned about size as relevant dimension in the animals task, should be able to apply this concept to a new set of stimuli (vehicles) of told that the same feature is relevant across domains. Moreover, if participants had learned more accurate representations of these rules under blocked training, they might perform better on this cross-domain transfer than those who had received interleaved training. To investigate this, we tested a new cohort of participants, with a separate group per combination of training domain (animals/vehicles) and curriculum (blocked/interleaved). All participants were only

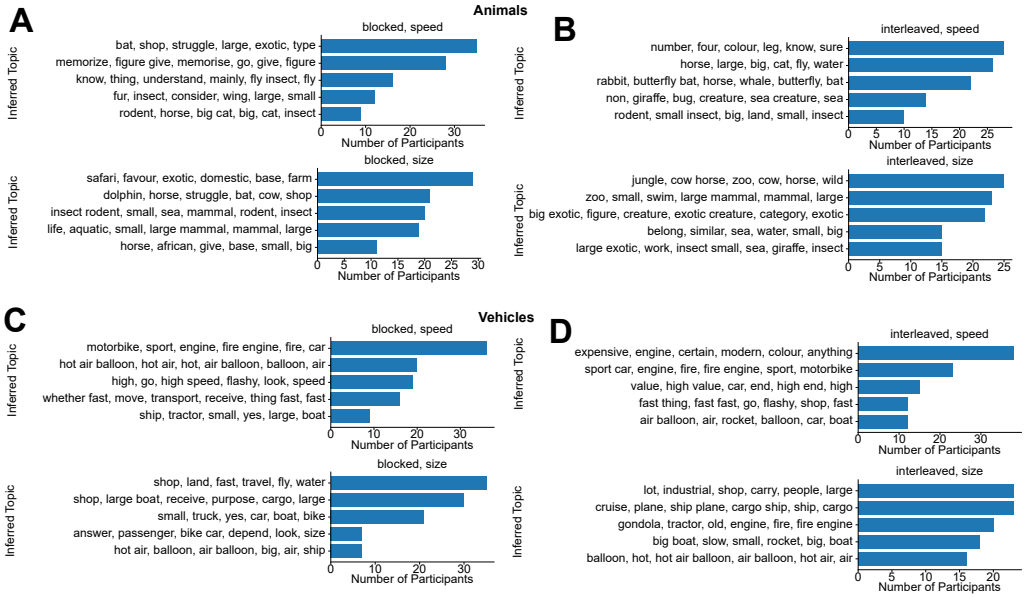

Figure 6.6: Topic Modelling, Experiment 1. Results from topic-modelling with non-negative matrix factorisation (NMF). Depicted is the frequency of the top five inferred topics per task, together with the top six words associated with each inferred topic. No salient differences between groups, but results indicate that many participants relied on alternative feature dimensions (habitat, value, exotic versus domestic, number of legs, could carry passengers or cargo) or individual exemplars.

trained on a single domain but evaluated on interleaved examples from the training domain (which we refer to as base domain or base tasks) and the other domain (transfer domain or transfer tasks) (**Fig. 6.7A**). Moreover, to gain more insights into the type of representation they had acquired, we asked all participants after the main experiment to arrange stimuli in a circular arena via drag and drop, so that distances between them corresponded to the rule they had learned for the orange or blue shop (i.e., cluster rewarding and non-rewarding stimuli together, or arrange them on a line based on their value) (**Fig. 6.7B**).

While the performance of all participants improved during training, reaching on average around 75% correct on a subsequent test phase, we again observed no difference in accuracy between the blocked and interleaved training groups (*animals, base:* $T(183) = 0.412 p = 0.681$, *vehicles, base:* $T(181) = 0.255 p = 0.799$, **Fig. 6.8A**). Performance on the transfer domain was above chance, but much worse than on the base task, with participants performing on average at 54% after training on animals, and 60% after training on the vehicles domain. While performance in our cohort was con-

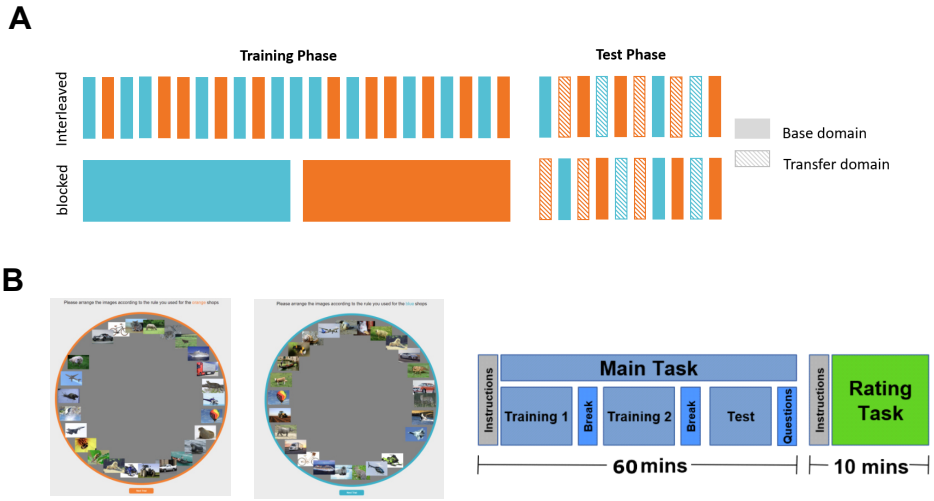

Figure 6.7: Task design, experiment 2. (**A**) In experiment two, participants were again either trained on the animals (experiment 2a) or vehicles domain (experiment 2b) and received either blocked or interleaved training. However, this time we tested them on interleaved trials both from the domain they had been trained on ("base") and the other domain ("transfer"), to assess cross-domain generalisation performance. (**B**) To gain more insights into the used strategies, we asked participants to arrange stimuli from both domains in a two-dimensional arena so that the distances between them corresponded to the rules/reward structures of the two shops (orange and blue). This arena task was completed by all participants after the main experiment.

sistently higher under blocked training, this difference was only significant for those trained on the animals domain (*animals, transfer:* $T(183) = 2.147 p = 0.033$, *vehicles, transfer:* $T(181) = 1.204 p = 0.230$, **Fig. 6.8A**). Replicating results from experiment 1, we observed no differences in the sensitivity of choices to the relevant dimensions between curricula (*animals, slope relevant:* $T(183) = 0.925, p = 0.356$, *offset relevant:* $T(183) = -1.824, p = 0.07$; *vehicles, slope relevant:* $T(181) = 0.136, p = 0.892$, *offset relevant:* $T(181) = -1.049, p = 0.296$, **Fig. 6.8B**), and intrusions from the irrelevant dimension were not larger under interleaved training (*animals, slope irrelevant:* $T(183) = 1.83, p = 0.069$, *offset irrelevant:* $T(183) = -1.6, p = 0.111$; *vehicles slope irrelevant:* $T(181) = 2.26, p = 0.025$, *offset irrelevant:* $T(181) = -0.53, p = 0.597$, **Fig. 6.8B**).

In experiment 1, we found some evidence for more accurate boundary estimates under blocked learning. Surprisingly, this effect did not replicate in experiment 2, and none of the parameters of the psychophysical model differed significantly between

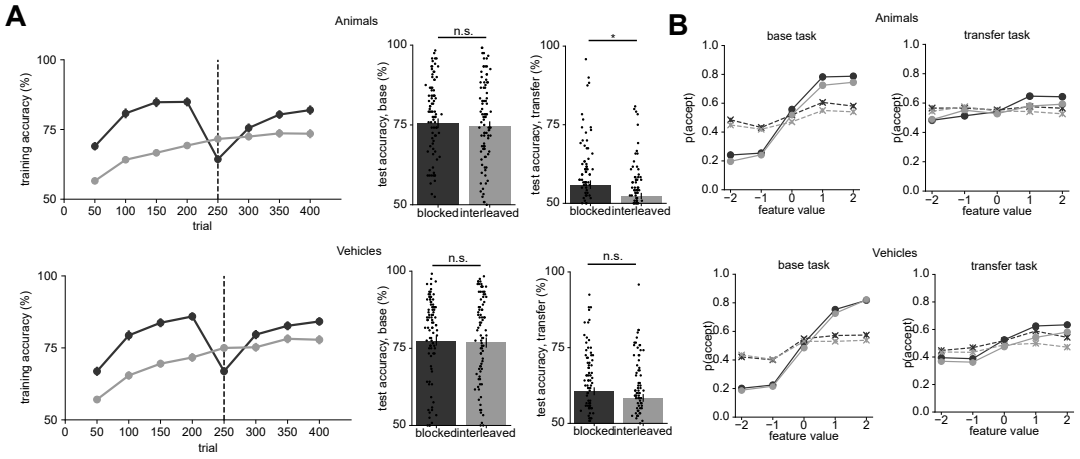

Figure 6.8: Behavioural results, experiment 2. (**A**) Learning curves and test phase accuracy on the base and transfer task. In both domains, performance improved over time course of training phase, but test phase accuracy on the base domain did not differ between groups. In participants trained on animals, however, performance on the transfer domain was slightly higher under blocked training. (**B**) Choice patterns as function of the relevant (solid) and irrelevant (dashed lines) feature dimensions, shown for the animals (top) and vehicles (bottom) domain. Participants were sensitive to the relevant dimension of the base domain (left), but much less so on the trials from the transfer domain (right). We observed no differences in sensitivity between blocked and interleaved groups, neither for the relevant nor the irrelevant dimensions.

blocked and interleaved training (*animals, baseline: angular bias* $T(183) = 0.698, p = 0.486$*, lapse* $T(183) = -0.364, p = 0.716$*, slope* $T(183) = 0.625, p = 0.533$*, offset* $T(183) = -1.879, p = 0.062$*; vehicles, baseline: angular bias* $T(181) = 1.963, p = 0.051$*, lapse* $T(181) = -0.574, p = 0.567$*, slope* $T(181) = 0.867, p = 0.387$*, offset* $T(181) = -0.933, p = 0.352$, **Fig. 6.9A-B**). While we had observed at least a small advantage of blocking over interleaving in transfer accuracy, neither the boundary estimate nor the lapse rate differed between these curricula (*animals, transfer: angular bias* $T(183) = -0.223, p = 0.824$*, lapse* $T(183) = -0.889, p = 0.375$*, slope* $T(183) = 0.132, p = 0.895$*, offset* $T(183) = 0.838, p = 0.403$*; vehicles, transfer: angular bias* $T(181) = 1.081$*, lapse* $T(181) = -0.657, p = 0.512$*, slope* $T(181) = 0.77, p = 0.438$*, offset* $T(181) = -0.986, p = 0.326$, **Fig. 6.10A-B**). How can we explain the worse performance on the generalisation task? One possibility is that they just learned associations between stimuli and responses. Such a rote-learning strategy would be

marked by error patterns in the transfer task that are unspecific and not dependent on the two feature dimensions. Indeed, across all domains and training curricula, we observed that lapse rates were significantly higher on the transfer compared to the base domain at test (*animals blocked* $T(90) = -9.179, p < 0.001$*, animals interleaved* $T(93) = -9.975, p < 0.001$*, vehicles blocked* $T(93) = -8.900, p < 0.001$*, vehicles interleaved* $T(88) = -7.459, p < 0.001$). We note, however, that lapse rates for the transfer task varied considerably within groups, which may suggest that strategies across the cohort were highly heterogenous.

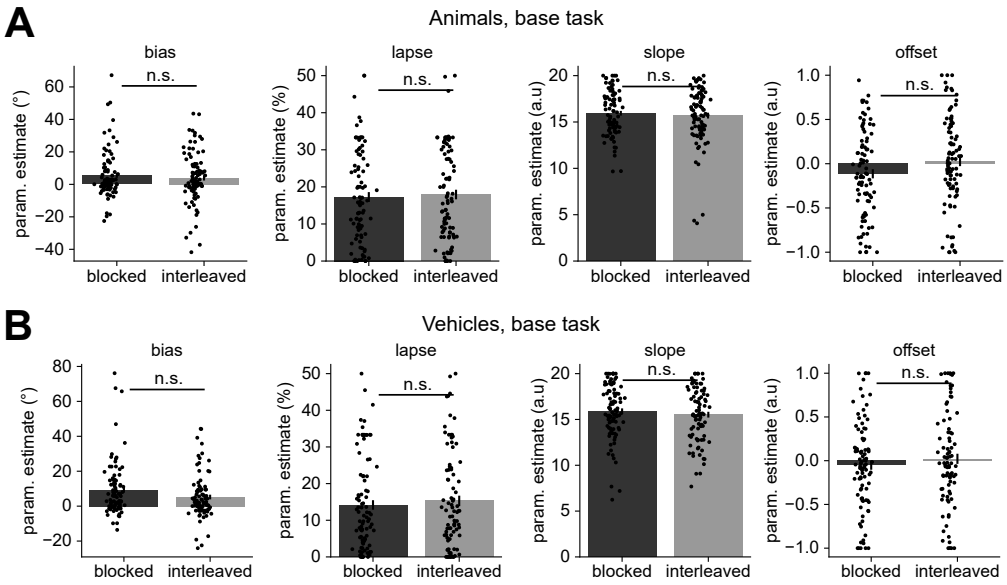

Figure 6.9: Psychophysical model fits, experiment 2, base domain (**A**) Parameter estimates for choice model fit to test trials from the base domain for those participants who were trained on animals. None of the parameter estimates differed significantly between the blocked and interleaved training groups. (**B**) Same as (A) but for vehicles.

6.2.4 Ratings provided in arena task are partially consistent with ground truth

The previous analyses provided no evidence for a benefit of blocking over interleaving, neither in performance on the base task, nor for cross-domain generalisation. One possibility might be that any effect present in the data was masked by the higher switch-cost demands of the test phase, which now included not only switches between con-

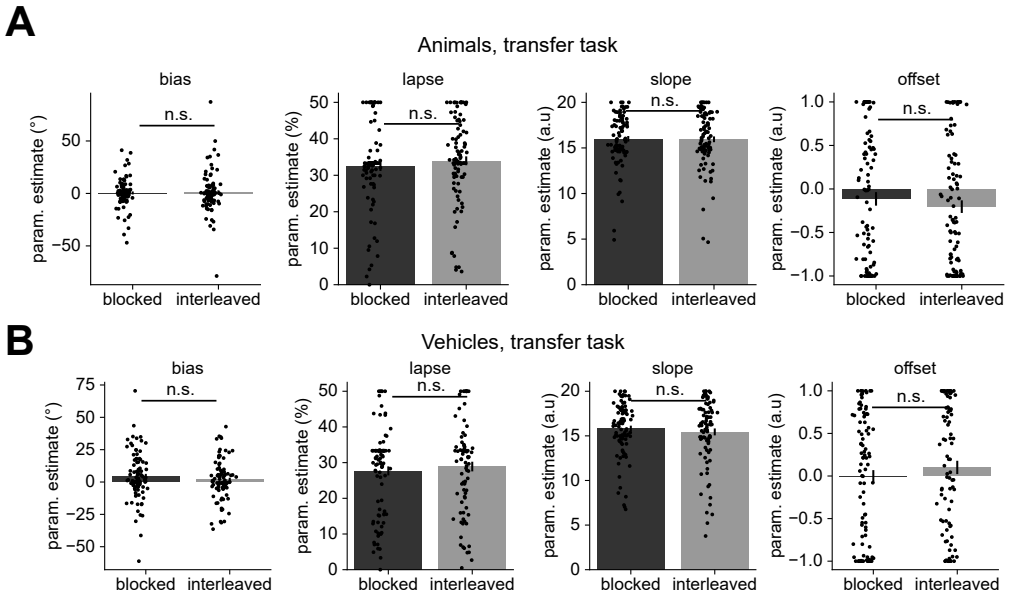

Figure 6.10: Psychophysical model fits, experiment 2, transfer domain (**A**) Parameter estimates for choice model fit to test trials from the transfer domain for those participants who were trained on animals. None of the parameter estimates differed significantly between the blocked and interleaved training groups. (**B**) Same as (A) but for those trained on vehicles.

texts, but also between domains. At the end of the main task, we asked our participants to arrange stimuli from both domains inside a circular arena, so that distances between stimuli corresponded to the rule they had learned for the orange and blue shops respectively. This allowed us to test not only whether they used rules that mapped onto the ground truth dimensions of size and speed, but also if ratings were consistent between the base and transfer domain. Averaging ratings of all participants using RSA-based techniques and visualising them in two dimensions provided some evidence that participants did indeed cluster by feature dimension, and that on average, these mapped onto the relevant dimensions of size and speed respectively). In other words, for the speed tasks, stimuli of both domains were grouped based on speed, while the size dimension was used to group stimuli in the size task (**Fig. 6.11A-B**). As in the previous study, we couldn not be certain whether participants were indeed aware of these variations along size and speed dimension or used other features to guide their choices. We could, however, test whether the alignment of these ratings with the ground truth dimensions differed between training groups. If there was a benefit of blocking on

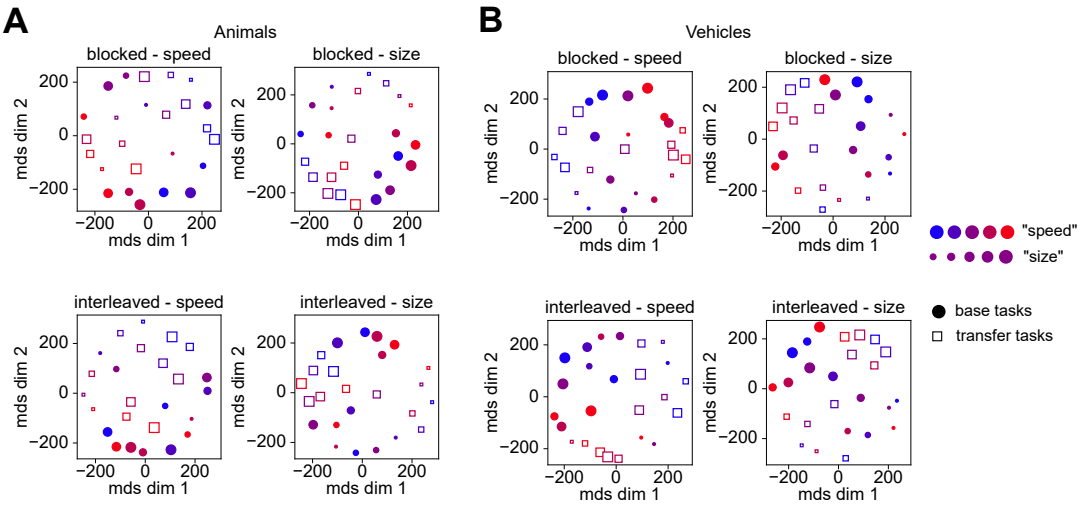

Figure 6.11: Visualisation of arena task results, experiment 2. (**A**) Visualisation of the group-level averages of arena task ratings for participants trained on the animals domain. Dots indicate stimuli from the base domain and rectangles stimuli from the transfer domain. Overall, base and stimuli appear to be arranged according to the same rule, with gradients of speed (indicated by dot/square colour) and size (indicated by dot/square radius/width) visible for the speed and size task respectively. No qualitative difference between blocked and interleaved training. (**B**) Same as (A) but for participants trained on the vehicles domain. Again, participants seemed to be able to arrange stimuli by size in the size task and – to a lesser extent – by speed in the speed task.

boundary learning, which was masked in the interleaved test phase but could be identified with the rating task, a model representing the task-relevant dimension should explain the ratings significantly better under blocked compared to interleaved training. To test this, we regressed the ratings against such a model, and control models encoding the irrelevant feature, the whole 5x5 grid or a projection onto the diagonal ranging from slow and small animals/vehicles to large and fast animals/vehicles. While overall, the relevant dimension model seemed to fit best, the difference between blocked and interleaved was only significant for the base task of the cohort trained on animals (*animals, base: blocked vs interleaved:* $T(183) = 2.061, p = 0.0407$*; vehicles, base: blocked vs interleaved:* $T(181) = 0.768, p = 0.4435$*; animals, transfer: blocked vs interleaved:* $T(183) = 0.671, p = 0.5029$*; vehicles, transfer: blocked vs interleaved:* $T(181) = 0.762, p = 0.4471$, **Fig. 6.12A-B**).

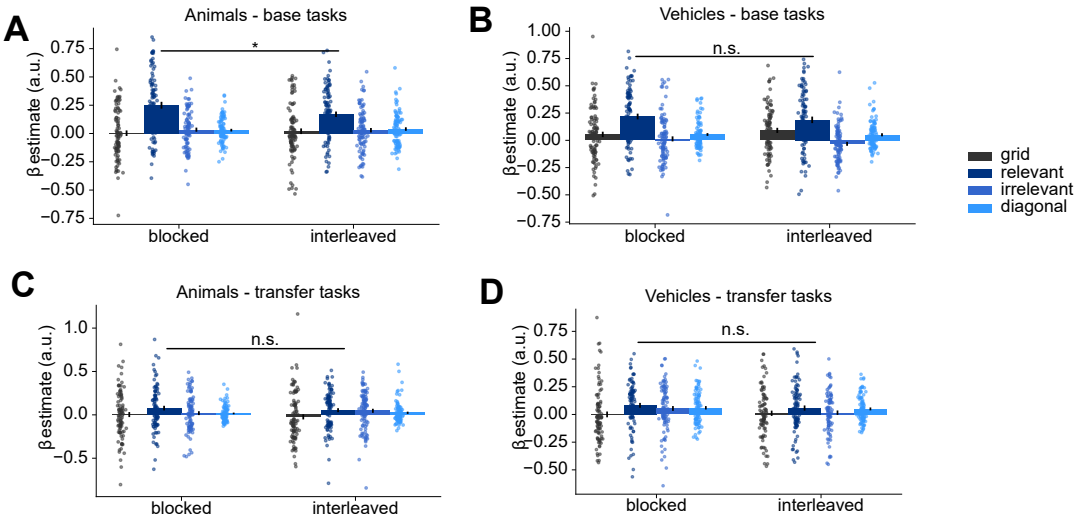

Figure 6.12: Model fits to arena task results, experiment 2. (**A-D**) Regression results for fits of grid, relevant, irrelevant dimension and diagonal model RDMs, fit to RDMs derived from the participant ratings. In both training domains and in the base as well as transfer task, the model encoding only the relevant dimension explains the ratings best. Differences in the fit between blocked and interleaved training, however, were only significant for the base task in those participants who had been trained on the animals dimension. Results suggest that on average, participants were able to arrange stimuli from the base and transfer domain according to the relevant rule in each task.

6.3 Discussion

The aim of this experimental series was two-fold. First, we tried to assess whether the benefit of blocking over interleaving which we had previously reported for context-dependent decision making would generalise to stimuli that varied along semantic dimensions. This prediction was not trivial, as the task posed challenges for learners that were not present in the original study. First, semantic stimuli could vary along many more dimensions, compared to stimuli that were parametrically varied along a hand-crafted set of perceptual dimensions. Secondly, participants could not infer these dimensions from variation in visual appearance alone but needed to rely on their semantic memory of features commonly associated with these stimuli. The second aim was to test whether participants could generalise their knowledge to another domain with the same task rules. Across two experiments with 400 participants each, we found only very limited evidence in support of these predictions, suggesting that further re-

finements of the experimental design are required.

In experiment 1, we trained and tested participants on context-dependent decisions within a single domain, where the reward depended either on the speed or size of stimuli. While performance improved over the time-course of the training session, test phase performance did not differ between learning curricula. We found some limited evidence for more accurate category boundaries under blocked training, but a benefit of this on accuracy was likely masked by other sources of error which our models failed to capture. Previous work suggested that processes underlying category learning might differ between so-called rule-based learning, in which single dimensions are relevant, and information-integration learning, for which multiple dimensions need to be integrated to make decisions (Ashby & Maddox, 2005). In our previous experiment, we had observed that in the latter, tested with diagonal category boundaries in a two-dimensional feature space, the benefit of blocking was masked by higher rates of non-specific errors (lapses) (Flesch et al., 2018). While technically, in the animals/vehicles task, only a single dimension was relevant, participants might have attempted to integrate both perceptual and mnemonic information to identify the correct response for each stimulus. However, at least with the psychophysical model we used to estimate these lapse rates, we could not identify any differences in these errors between groups. Topic modelling of written reports provided by the participants after the main experiment indicated that, while a large fraction correctly identified speed or size, others relied on dimensions we had not anticipated, such as the habitat of animals, or the monetary value of vehicles. Moreover, this exploratory analysis revealed that many participants tried to memorise the mappings of individual animals to responses, rather than learning the latent rule.

A simple explanation for the absence of differences in experiment 1 would be that it did not sufficiently motivate participants to learn the correct dimensions, as rote-learning of mappings between animal/vehicle types and their associated responses was sufficient to perform well on the training and test phase. This was different in experiment 2, where on average, participants performed above chance on the transfer domain, with a slight benefit of blocked over interleaved training. However, lapse rates

were substantially higher in all groups on transfer compared to base domain trials, indicating that at least a large fraction of participants failed to generalise the correct rules across domains, or that this was overshadowed by other non-specific sources of error, such as higher switch costs compared to experiment 1. While we did not test for this explicitly, data from the rating task hinted at least at partial agreement between the rules participants used for the base and transfer task, but no consistently stronger alignment with the ground truth under blocked compared to interleaved training.

A key take-away from these findings is that changes in experimental design are required to test for the differences in the effectiveness of learning curricula in abstract spaces. In our previous study, we reported that the blocking benefit was particularly pronounced for participants who had a strong prior on the two task-relevant dimensions (Flesch et al., 2018). In other words, those who realised that the stimuli varied along two feature dimensions found it easier to map the context-specific rules onto these dimensions, and even more so if they were given a blocked training curriculum. To test for a similar benefit in semantic spaces, we would need to alter the experimental design in a way that it induces priors on the relevant dimensions. But how can we provide this information to participants without trivialising the problem? Recent evidence suggests that humans can be trained on transitivity problems with arbitrary mappings between images and words that don't carry any specific meaning. Nelli et al., (2021) trained participants on the "brispiness" of novel, previously unseen computer-generated images and later tested their ability to perform transitive judgements based on this concept of brispiness. Similarly, we could train participants on pairs of animals/vehicles and ask which of the two has a larger value, and later on assess their ability to map these pairwise comparisons onto the whole feature space. This could serve as a pre-training exercise that alerts our participants of potentially task-relevant dimensions, without giving them explicit information about the relevance of size and speed in particular.

Leaving these limitations of the design aside, there are several clear avenues for future research. In our experiment, we asked whether participants would map information on domain-specific or domain-general representations of size and speed. We explicitly told participants that stimuli from the transfer domain would map onto the

same features as stimuli from the base domain. It remains unclear if and how humans could learn about the domain generality of concepts without explicit instruction. In the previous chapter, we motivated the idea that humans may have strong priors on contexts being stationary, which we modelled with a "sluggishness" parameter that led to interference between distinct tasks under interleaved training. It may, however, be possible that this sluggishness is beneficial in cases where contexts are highly similar, as it promotes the formation of a joint task representation. One could test whether domain-general representations are learned better when similar rules from different domains are interleaved during training, as in a curriculum in which participants are first trained on size in animals and vehicles, followed by speed in these two domains. Early work on category learning supports this hypothesis, as for example Carvalho & Goldstone, (2014) demonstrated that blocking confers a benefit when examples within categories are highly variable, as it highlights which features are shared within a block.

Another avenue for future research is the neural geometry that may accompany cross-domain transfer. Most of the work described in this thesis used orthogonal rules which were mapped onto orthogonal task representations, to minimise interference between tasks. In the context of the current study, we would expect to find that size and speed are again mapped onto orthogonal coding axes, both within and across domains (e.g., size of animals versus speed of vehicles). We would hypothesise that a neural code which optimises for the transfer demands of the tasks, however, would then map representations of the same rule across domains on parallel planes (e.g., size in animals versus size in vehicles), as the same readout could be applied zero-shot to the base and transfer domain. Indeed, in a pilot simulation with a neural network trained on the animals/vehicles task with interleaved data, we found some preliminary evidence for representations consistent with this prediction (**Appendix C, Supplementary Methods and Fig. C.4**). Further evidence for this prediction comes from a recent study in which macaques were trained on a context-dependent decision problem that involved either holding or releasing a lever in response to various fractal images. Here, the authors found strong evidence for parallel representations, in which the rules generalised across contexts, in prefrontal regions (Bernardi et al., 2020). Similarly, a recent EEG

study with humans found that when trained on a magnitude comparison task (more or less judgements) with different numerical ranges experienced in different contexts, both humans and recurrent neural networks embedded this context-specific knowledge on parallel planes, which enabled cross-context generalisation of the concept of magnitude (Sheahan et al., 2021).

However, testing these predictions in human participants requires an experimental paradigm that promotes learning of the ground-truth semantic rules intended by the experimenter, rather than relying on rote-learning strategies or other feature dimensions. While further refinements of this paradigm were beyond the scope of this doctoral thesis, work on the project is still ongoing.

6.4 Methods

6.4.1 Software

The behavioural experiments were implemented in JavaScript, using a lab-internal toolbox developed for running online studies. All statistical analyses were carried out in Python with the Pandas 1.2.3, NumPy 1.19, SciPy 1.60, Statsmodels 0.13 and Scikit-Learn 0.24.1 packages. Topic modelling was carried out with the WordCloud 1.8.1 and nltk 3.7 packages. Figures were generated with Matplotlib 3.3.2.

6.4.2 Participants

A total of 400 participants (100 per combination of domain (animals/vehicles) and training curriculum (blocked/interleaved), 264 female, 130 male, 4 other, 1 preferred not to say, mean age 26.92) were recruited for experiment 1, using the Prolific crowdsourcing platform. The same number of participants was recruited for experiment 2, again on Prolific, but data from four participants was lost due to technical issues on the server (259 female, 132 male, 4 other, mean age 28.89). Participants were compensated for their time at a rate of £10/hour. We restricted recruitment to participants in the age range 18-40 who were ordinarily residents in the UK and had an approval rating of at least 85% averaged over their past five submissions. All experiments were approved by the Medical Sciences Research Ethics Committee of the University of Oxford (approval reference: R50750/RE001). Participants with training performance

at or below chance (50%) or who had missed more than 25% of test trials were excluded from data analysis, leaving 89 for animals-blocked, 88 for animals-interleaved, 93 for vehicles-blocked and 90 for vehicles-blocked in the first experiment, and 90 for animals-blocked, 93 for animals-interleaved, 93 for vehicles-blocked and 88 for vehicles-blocked in the second experiment.

6.4.3 Task Design

Experiment 1 and 2 were run in forced-fullscreen mode. In both experiments, participants were instructed to learn about the customer preferences in two different shops (called the orange and blue shop) via trial-wise feedback. For each experiment, there were four different participant groups in total, one for each combination of training domain (animals/vehicles) and training curriculum (blocked/interleaved). In experiment 1, participants were trained and tested on the same domain (animals in 1a, vehicles in 1b). In experiment 2, participants were tested on the domain they had been trained on and the held-out domain, which we refer to as "base task" and "transfer task" respectively (2a- trained on animals, 2b – trained on vehicles). On each trial across all experiments, participants would either accept or reject an animal/vehicle and receive a reward based on how likely the chosen animal/vehicle was to sell in this store. Participants were not alerted to the semantic dimensions (size, speed) a priori but told that different features would be relevant in each shop, and that those features would scale with the reward received for accepting the animal/vehicle, with negative rewards indicating that it wouldn't sell well, and positive rewards that it would be liked by the customers of that shop. In both experiments, participants were first trained with feedback, either on a blocked or interleaved curriculum (200 trials per task), and performance was subsequently evaluated in an interleaved test phase without feedback (200 test trials in experiment 1 and 300 test trials in experiment 2). In the blocked curriculum, participants learned first about one task and then about the other (one block of 200 trials per shop), while in the interleaved curriculum, all 400 training trials were randomly shuffled. We equated the number of trials for each condition (size (5) x speed (5) x context (2)) and showed trial-unique stimuli in the training and test phase. In both domains, shops were either orange or blue, but differed in their appearance otherwise,

so that participant in experiment 2 could easily distinguish between the blue animal and the blue vehicle shop, for instance. For experiment 2, we told participants that customers with similar preferences would go to the two transfer shops, meaning that the features that determined rewards in the two shops of the base task (for example with animals) mapped onto the respective shop with the same colour in the transfer task (for example with vehicles).

In both phases, trials began with the display of the contextual cue (orange or blue shop) which remained on the screen throughout the trial. After a short interval (1000ms), the context was blurred and a stimulus was shown on the screen, together with the response contingencies for this trial. Participants used the f and j keys on their keyboard to either accept or reject a stimulus. The mapping was randomised across trials and indicated via boxes with the words "accept" and "reject", shown on either side of the stimulus. For example, if "accept" was shown left and "reject" right, participants used f to accept and j to reject a stimulus. Responses were possible for up to 300ms after stimulus onset, after which the trial timed out and the stimulus was automatically rejected. The chosen option was highlighted by a black rectangle drawn either around "accept" or "reject", which stayed on the screen for 500ms. In the training phase, participants received numerical feedback for accepting an animal, which scaled with the feature value along the relevant dimension within each shop, and always zero for rejecting a stimulus. The numerical feedback was shown for the chosen and unchosen option and replaced the "accept" and "reject" response contingencies once a button was pressed. The chosen option was highlighted by a black rectangle. Unbeknownst to the participants, the stimuli varied in five discrete steps along the size and speed dimension. In each shop, only one dimension was relevant, for example size in the orange store and speed in the blue store. The five levels of size/speed were mapped onto numerical rewards ranging from -50 to +50 in five steps. The feedback was displayed for 1000ms. Trials were separated by an ITI of 1000ms. No feedback was given during the test phase.

After successful completion of the main experiment, we asked participants to write down the rules they had identified in two text boxes, one per shop. Only in ex-

periment 2, participants then completed a short stimulus rating experiment (see below) in which they arranged stimuli from both domains via drag and drop in a circular arena according to the rules that applied to each shop (one arena per shop). To assess how participants represented the rules they had learned for the base and transfer task in experiment 2, we asked them to complete a rating task at the end of the main experiment. Participants had to complete four trials in total, two per shop. On each trial, 12/13 stimuli from both domains were interdigitated, spanning the whole 5x5 grid of size and speed (for example, mapping animals on odd linear indices and vehicles on even linear indices, with the opposite assignment on the subsequent trial). These were displayed on random locations around a circular aperture, whose edge colour indicated whether the trial belonged to the orange or the blue shop. Participants were asked to re-arrange these stimuli inside the circular "arena" so that the distance between stimuli depended on the reward associated with accepting them. Had they perfectly learned the rules and understood that the base and transfer stimuli vary along the same dimensions, they had arranged them on a line with five clusters, one per reward level (-50 to +50). We chose to interdigitate stimuli from both domains so that participants could express the relationship between the two on a single trial.

Note that initially, we ran two versions of experiment 2, with and without randomly allocated no-feedback trials during the training phase (9 out of 25 conditions per task). The purpose of these trials without feedback was to encourage participants to learn abstract rules, rather than individual stimulus-response mappings. We tested the same number of participants in both versions (200 per version). However, we observed no difference between the version with and without partial feedback. For simplicity, in this report, we decided to collapse over the two datasets, which yielded sample sizes similar to those reported for experiment 1 (400 in total, see participant info above).

6.4.4 Quantification and statistical analyses

Accuracy and human choice patterns: Analysis of accuracy and choice patterns followed the procedures described in detail in chapter 4 (human fMRI study). Again, we quantified choice patterns using (a) sigmoidal curves fit separately to the relevant and

irrelevant dimensions, (b) a linear model in which we regressed RDMs derived from test phase choice patterns on linear/factorised model RDMs and (c) a psychophysical model with free parameters for the boundary error and lapse rate, slope and offset of a sigmoidal transducer.

Word clouds and topic modelling: To gain further insights into the rules participants used to solve the tasks, we used topic modelling techniques to analyse the written reports submitted at the end of the experiment. Participants were asked to report in two text boxes, one per shop, which rule they applied to decide whether to accept a stimulus in a shop. We first pre-processed the written responses at single subject level using a standard pipeline for natural language processing. First, we removed leading/trailing whitespace and escape characters and transformed all words to lower case. Next, we lemmatized the words using functions from the NLTK package. Lastly, we removed special characters and tokenised the words using the *wordtokenize* function from the nltk package. As first exploratory analysis, we then calculated word clouds, which show the n most frequent words, where the fontsize scales with the frequency of the word in the text. We performed this analysis separately for each participant group and shop (e.g. orange shop with animals), using the wordcloud Python package. Next, we turned to a technique from natural language processing called topic modelling, to reveal clusters of rules reported by our participants (so-called topics). More specifically, we applied non-negative matrix factorisation (NMF) to the reports, which decomposes a word-by-report matrix into a word-by-topic and a topic-by-document matrix. In other words, it groups the reports by latent variables, the inferred "topics". The first matrix in this decomposition ascribes weights to words, denoting for each word how likely it is to occur in a given topic, while the second matrix assigns these topics to individual documents. To understand which rules our participants reported and if this depended on the domain or curriculum, we applied this technique separately to reports from each group and task. We then identified the top five topics for each group and counted the number of participants whose reports matched these topics. These histograms were visualised together with the top six words per topic.

Arena rating task: Subjective ratings from the arena task were analysed using an RSA-based approach. We first computed RDMs encoding the pairwise Euclidean distances between all stimulus coordinates at single trial level, yielding a single 25x25 RDM per participant and trial. For visualisation purposes, we averaged over the normalised single-subject RDMs and projected them into two dimensions using multidimensional scaling. Next, we regressed these RDMs against two model RDMs encoding the task-relevant or task-irrelevant feature, a control model that represented a task-agnostic 5x5 grid and a diagonal model in which stimuli were mapped onto the main diagonal (from low size and speed to high size and speed. We fit these at single-subject level using linear regression and performed unpaired t-tests on the parameter estimates to test for differences between training groups (blocked versus interleaved) within each training domain.

General Conclusion

The aim of this thesis was to further our understanding of the computations and representations underlying continual task performance. I focussed specifically on context-dependent decision-making problems in which different features of the same type of stimulus are relevant depending on the context in which it occurs. As each chapter ended already with a detailed discussion, here I will only briefly revisit the key findings and implications, relate the results from the individual chapters to each other and discuss limitations as well as potential future lines of research.

In Chapter 1 and 2, I reviewed the literature on cognitive control and discussed the utility of artificial neural networks as models of representation learning. In **Chapter 1** I argued that the most impactful theories on cognitive control proposed that PFC implements a task-specific gating strategy, so that irrelevant information is wholly or partially filtered out in order to avoid interference (Miller & Cohen, 2001). Studies on the representational geometry of neural codes in PFC, however, have thus far provided conflicting evidence as to whether the brain represents information in such a task-specific or instead in task-agnostic format. Previous empirical work had predominantly focussed on non-human primates, and it remained unclear how these representations could be learned in the first place. In **Chapter 2**, I introduced deep artificial neural networks as toolkit to study how representations are sculpted by task demands and argued that one should start with small tractable models, before studying representational commonalities between complex architectures and various brain regions.

Chapter 3 introduced most of the theoretical groundwork. Using a context-dependent

decision-making paradigm, I studied how changes in the scale of weights at initialisation affect how and what a neural network learns. I identified two different solutions for this problem, which we call "rich" and "lazy" regime, borrowing nomenclature from deep learning theory (Chizat et al., 2020; Woodworth et al., 2020). In both regimes, the networks were able to learn context-dependent responses, but the learning dynamics and acquired representations differed substantially. Solutions found under lazy learning were acquired fast, but less robust to noisy perturbations. Internal representations were task-agnostic, and RSA on the hidden layer revealed that the network had mostly recapitulated the structure present in the input space, so that learning was confined to the readout weights. In contrast, rich learning converged considerably slower, but the solution was much more tolerant to the addition of input noise and accompanied by highly structured representations in the hidden layers. More specifically, we observed orthogonal, low-dimensional manifolds, in which only task-relevant information was maintained and projected onto a separate axis in coding space. We then introduced the theory of non-linear gating, inspired by theories of cognitive control, which explains how these representations are learned in the rich regime. Complementary simulations with deeper neural networks revealed that under rich learning, representations might be progressively transformed along the layer hierarchy, so that irrelevant information is first filtered out, before task information is then projected into the frame of reference of the response. The challenge to represent multiple tasks with minimal interference might promote highly task-specific, orthogonal and low-dimensional representations. Interestingly, however, we noticed that a network trained in the lazy regime was also able to solve the task, and still generalise fairly well to previously unseen exemplars.

In **Chapter 4**, we investigated whether the human brain solves a context-dependent decision-making problem with task-specific or task-agnostic representations. Participants learned to either accept or reject high-dimensional stimuli that varied parametrically along two dimensions, of which only one was relevant in each context. We observed that after prolonged training on a blocked curriculum, which we have previously found to promote optimal task performance (Flesch et al., 2018), representations in fronto-parietal areas of the brain were highly structured, and exhibited a geome-

try consistent with the one predicted by rich learning. In contrast, representations in early visual cortex were mostly task-agnostic and described best by a grid-like model that represented the two features of the stimulus space along two different axes. In a supplementary re-analysis of a freely available dataset with recordings from FEF in macaques who had been trained on a comparable task, we sought to test more fine-grained predictions from the gating theory. While overall, the results seemed to be consisted with our predictions, we note that the way in which responses were counter-balanced differed between our fMRI study and the NHP experiment, which suggests that further experiments would be required to support our observation that effects fully generalise across species and recording modalities. Broadly speaking, however, the evidence provides novel insights into the coding scheme utilised by the brain to solve context-dependent decision problems that are learned from trial-wise feedback, and is consistent with the solutions learned under rich learning presented in the previous chapter

In **Chapter 5**, we proposed a model of human continual learning. In previous work, we had reported that humans benefit from training curricula in which they learn one task at a time, compared to a fully interleaved curriculum (Flesch et al., 2018). While results from the preceding chapters suggested that the brain learns factorised representations, and our modelling work demonstrated that these arise in artificial neural networks under rich learning, the training curricula used for human participants and neural networks were not comparable, as the latter had to be trained on interleaved data to prevent catastrophic forgetting. Harnessing insights from chapter 3, we first tested whether non-linear gating would in principle be sufficient to prevent forgetting. Next, we introduced an algorithmic motif, a simple Hebbian update step, which allows the network to learn such a gating scheme online, without further intervention by the experimenter. Lastly, we introduced the concept of "sluggish" task units, which maintain context information over successive trials, to successfully model the cost of interleaved training. As the model was explicitly developed to capture the error patterns in the two-context case, further work is needed to understand the mechanism underlying lifelong, continual learning, and how the proposed orthogonalization might

interact with replay during sleep. A key takeaway from our simulations, however, is that we stipulate that the brain uses time itself as a cue to form task sets, and that this prior on contexts being stationary introduces a cost in settings where context switches occur more frequently, such as in controlled psychological experiments with interleaved trials.

Finally, in **Chapter 6**, we attempted to study whether the benefit of blocked over interleaved learning for context-dependent decision problems generalises to abstract domains and promotes cross-domain transfer. The latter was motivated by a series of recent findings which indicated that the brain uses abstract representations to encode domain-general information, and that this information lies on parallel, rather than orthogonal axes, permitting generalisation across contexts (Bernardi et al., 2020; Sheahan et al., 2021). Further support for this hypothesis was provided by results from our simulations in earlier chapters, which had revealed parallel representations in deeper layers, which coded for generalisable information such as the magnitude of task-relevant features. The first experiment replicated our previous results in part, as the category boundary estimate of participants trained on blocked data was more closely aligned with the ground truth. In the second experiment, we found some very limited evidence for a benefit of blocked over interleaved training on performance on the transfer domain, but results from experiment 1 did not replicate. Analyses of error patterns and debriefing reports revealed that many participants had failed to identify the correct semantic dimensions, and either relied on rote-learning strategies or assumed that other semantic dimensions which we had not considered were task relevant. Hence, we could unfortunately not say with certainty which training regime might promote generalisation in semantic spaces and provided some suggestions for follow-up experiments in which participants first learn along which semantic dimensions the stimulus sets vary, which will hopefully reduce the hypothesis space and resolve the issue with rote learning.

Across four chapters, we provide converging evidence that the brain relies on highly task-specific neural codes to solve the challenge of context-dependent decision mak-

ing. Mechanistically, we propose that these representations are implemented by a top-down gating signal, which silences neurons encoding task-irrelevant dimensions and increases the gain of neurons that represent information relevant to the current task goal. As this information was partitioned into separate populations of units, we observed axis-aligned, orthogonal representations at the population level. Furthermore, we demonstrate that this gating scheme may provide a solution for continual learning, and demonstrate how Hebbian learning mechanisms can be used to partition knowledge via gating signals continually.

The thesis provides further evidence in favour of early theories of cognitive control, which have postulated that prefrontal cortex gates relevant information in a context-dependent manner (Miller & Cohen 2001). Furthermore, our investigations of neural geometries demonstrate that earlier reports of NHPs trained on comparable tasks (Roy et al., 2010) might generalise to human participants. Previous theoretical work has explored the utility of these gating signals for continual learning (Masse et al., 2018; Russin et al., 2022). In Chapter 5, we demonstrate that a simple Hebbian mechanism might be sufficient to implement this gating strategy.

It shall be noted, however, that this thesis focussed on a very narrowly defined perspective on representation learning, which leaves ample opportunity for follow-up work. First, our work did not address the question of how the brain learns to represent information in the first place. While the neural networks were trained on the tasks tabula-rasa, our adult human participants were able to capitalise on rich priors, which have slowly matured over the timespan of several decades in which they were exposed to changing situational demands. Specifically, I argued that top-down gating mechanisms act on a population of neurons with highly axis-aligned codes, in which various features are mapped onto distinct neural sub-populations. Related work suggests that the brain forms a basis set of low-dimensional, abstract codes (such as for magnitude) which can be used as scaffold for various task-specific demands (Summerfield et al., 2020). The precise conditions that promote such a mapping, as well as the learning

rules employed by the brain, however, are topics for future research.

On a related note, this work focussed on the special case of a two-dimensional feature space with orthogonal task rules. Whether the orthogonal representations in fronto-parietal areas were a mere by-product of this organisation, or if the brain actively orthogonalizes representations to minimise interference, is an open research question. Some work from the domain of working memory, however, does indeed point at the existence of such an active orthogonalization mechanism (Panichello & Buschman, 2021). Similarly, one might wonder how a fundamentally new concept could be learned and integrated into the extant neural code, without disrupting past knowledge (Nelli et al., 2021). Some preliminary evidence indicates that neural representations "drift" over time, with selectivity to specific features slowly moving across a neural population (Rule et al., 2019). Future lines of research could investigate these questions with longitudinal studies, that track the evolution of neural representations from the first exposure to a new concept until skill mastery. One might also ask if these long-term representational changes are restricted to specific age groups or appear in children and adults alike.

Furthermore, whether the brain uses a task-specific or task-agnostic code might depend on the task demands or the specifics of the training curriculum. Here we subjected our participants to an interleaved test phase, which introduced a considerable switch cost. It could be the case that the brain compressed the irrelevant dimension to a lesser extent in blocked test phases, where less trial-to-trial interference was present. This relates to the broader distinction between automaticity and control, and how this trade-off maps onto neural representations. In related work on multi-tasking, it has been suggested that there is a fundamental trade-off between the ease with which new tasks can be acquired, facilitated by shared representations, and the mitigation of conflict to promote multitasking capability, which requires more separated representations (Musslick & Cohen, 2021).

Interestingly, we found evidence for orthogonal representations across the whole fronto-parietal network. As mentioned in the introduction, earlier studies on neural correlates of cognitive control attributed different functional roles to subregions within

this network, such as error monitoring in ACC (Kerns et al., 2004). We found some weak evidence for correlations between the orthogonality of representations and the precision of behavioural responses. Future studies could investigate how these representations facilitate various constituent processes of cognitive control, and derive a more fine-grained understanding of the postulated role of orthogonal representations in minimising cross-task interference.

The goal of this doctoral work was to contribute to our understanding of the computations and representations underlying continual task performance. Together, the presented work provides further support for the theory that the prefrontal cortex gates information in a task-specific manner, presumably in order to avoid interference between contexts. Our computational simulations suggests that the brain could use Hebbian learning mechanisms to acquire these gating schemes. The implications are not restricted to neural theory but do also shed some light on potential pitfalls that could occur when representations of off-the-shelf deep neural networks are compared to those in the brain. It highlights the utility of small, tractable classes of models as starting point to further our understanding of neural computation. While my doctoral work has only scratched the surface, I hope that it has contributed to our understanding of representation learning and might potentially spark inspiration for future projects that apply insights from machine learning theory to the study of the human brain.

Bibliography

Akhtar, N. and Mian, A. (2018). Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access*, 6:14410–14430.

Aoi, M. C., Mante, V., and Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. *Nature Neuroscience*, 23(11):1410–1420.

Aron, A. R., Monsell, S., Sahakian, B. J., and Robbins, T. W. (2004). A componential analysis of task-switching deficits associated with lesions of left and right frontal cortex. *Brain*, 127(7):1561–1573.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. *arXiv:1901.08584 [cs, stat]*.

Asaad, W. F., Rainer, G., and Miller, E. K. (2000). Task-Specific Neural Activity in the Primate Prefrontal Cortex. *Journal of Neurophysiology*, 84(1):451–459.

Ashby, F. G. and Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology*, 56:149–178.

Badre, D., Bhandari, A., Keglovits, H., and Kikumoto, A. (2021). The dimensionality of neural representations for control. *Current Opinion in Behavioral Sciences*, 38:20–28.

Baker, N., Lu, H., Erlikhman, G., and Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, 14(12):e1006613.

Barak, O., Rigotti, M., and Fusi, S. (2013). The Sparseness of Mixed Selectivity Neurons Controls the Generalization–Discrimination Trade-Off. *Journal of Neuroscience*, 33(9):3844–3856.

Baram, A. B., Muller, T. H., Nili, H., Garvert, M. M., and Behrens, T. E. J. (2021). Entorhinal and ventromedial prefrontal cortices abstract and generalize the structure of reinforcement learning problems. *Neuron*, 109(4):713–723.

Barnes, J. M. and Underwood, B. J. (1959). "Fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58(2):97–105.

Behrens, T. E. J., Muller, T. H., Whittington, J. C. R., Mark, S., Baram, A. B., Stachenfeld, K. L., and Kurth-Nelson, Z. (2018). What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron*, 100(2):490–509.

Benna, M. K. and Fusi, S. (2016). Computational principles of synaptic memory consolidation. *Nature Neuroscience*, 19(12):1697–1706.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4):954–967.

Boorman, E. D., Behrens, T. E., and Rushworth, M. F. (2011). Counterfactual Choice and Learning in a Neural Network Centered on Human Lateral Frontopolar Cortex. *PLOS Biology*, 9(6).

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., and Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F. G., Hummel, J., Heaton, R. F., Evans, B. D., Mitchell, J., and Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *PsyArXiv*.

Brincat, S. L., Siegel, M., von Nicolai, C., and Miller, E. K. (2018). Gradual progression from sensory to task-related processing in cerebral cortex. *Proceedings of*

the National Academy of Sciences of the United States of America, 115(30):E7202–E7211.

Brown, R. E. (2016). Hebb and Cattell: The Genesis of the Theory of Fluid and Crystallized Intelligence. *Frontiers in Human Neuroscience*, 10:606.

Buchsbaum, B. R., Greer, S., Chang, W.-L., and Berman, K. F. (2005). Meta-analysis of neuroimaging studies of the Wisconsin card-sorting task and component processes. *Human Brain Mapping*, 25(1):35–45.

Carey, S. (2011). Précis of The Origin of Concepts. *Behavioral and Brain Sciences*, 34(3):113–124.

Carvalho, P. F. and Goldstone, R. L. (2014). Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & Cognition*, 42(3):481–495.

Carvalho, P. F. and Goldstone, R. L. (2015). What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, 6:505.

Chaisangmongkon, W., Swaminathan, S. K., Freedman, D. J., and Wang, X.-J. (2017). Computing by Robust Transience: How the Fronto-Parietal Network Performs Sequential, Category-Based Decisions. *Neuron*, 93(6):1504–1517.

Chaudhry, A., Khan, N., Dokania, P. K., and Torr, P. H. S. (2020). Continual Learning in Low-rank Orthogonal Subspaces. *arXiv:2010.11635*.

Chaudhuri, R., Gerçek, B., Pandey, B., Peyrache, A., and 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.

Chizat, L., Oyallon, E., and Bach, F. (2020). On Lazy Training in Differentiable Programming. *arXiv:1812.07956 [cs, math]*.

Cho, R. Y., Nystrom, L. E., Brown, E. T., Jones, A. D., Braver, T. S., Holmes, P. J., and Cohen, J. D. (2002). Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. *Cognitive, Affective, & Behavioral Neuroscience*, 2(4):283–299.

Cohen, J. D., Dunbar, K., and McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review*, 97(3):332–361.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., Linenweber, M. R., Petersen, S. E., Raichle, M. E., Van Essen, D. C., and Shulman, G. L. (1998). A Common Network of Functional Areas for Attention and Eye Movements. *Neuron*, 21(4):761–773.

Cueva, C. J., Saez, A., Marcos, E., Genovesio, A., Jazayeri, M., Romo, R., Salzman, C. D., Shadlen, M. N., and Fusi, S. (2020). Low-dimensional dynamics for working memory and time encoding. *Proceedings of the National Academy of Sciences*, 117(37):23021–23032.

Çukur, T., Nishimoto, S., Huth, A. G., and Gallant, J. L. (2013). Attention during natural vision warps semantic representation across the human brain. *Nature Neuroscience*, 16(6):763–770.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314.

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

Dujmović, M., Malhotra, G., and Bowers, J. S. (2020). What do adversarial images tell us about human vision? *eLife*, 9:e55978.

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

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

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

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

Farajtabar, M., Azizan, N., Mott, A., and Li, A. (2019). Orthogonal Gradient Descent for Continual Learning. *arXiv:1910.07104 [cs, stat]*.

Ferguson, G. A. (1956). On transfer and the abilities of man. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 10(3):121–131.

Flesch, T., Balaguer, J., Dekker, R., Nili, H., and Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, 115(44):E10313–E10322.

Flesch, T., Juechems, K., Dumbalska, T., Saxe, A., and Summerfield, C. (2022). Orthogonal representations for robust context-dependent task performance in brains and neural networks. *Neuron*, 110(7):1258–1270.e11.

Flesch, T., Nagy, D., Saxe, A., and Summerfield, C. (2021). Modelling continual learning in humans with Hebbian context gating. In *Cosyne Abstracts*.

Franklin, N. T. and Frank, M. J. (2020). Generalizing to generalize: Humans flexibly switch between compositional and conjunctive structures during reinforcement learning. *PLOS Computational Biology*, 16(4):e1007720.

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

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

Fusi, S., Miller, E. K., and Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. *Current Opinion in Neurobiology*, 37:66–74.

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

Gao, P. and Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, 32:148–155.

Gao, P., Trautmann, E., Yu, B., Santhanam, G., Ryu, S., Shenoy, K., and Ganguli, S. (2017). A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv*, page 214262.

Geiger, M., Jacot, A., Spigler, S., Gabriel, F., Sagun, L., d'Ascoli, S., Biroli, G., Hongler, C., and Wyart, M. (2020). Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231 [cs, q-bio, stat]*.

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., and Wichmann, F. A. (2020). Generalisation in humans and deep neural networks. *arXiv:1808.08750 [cs, q-bio, stat]*.

Gisiger, T. and Boukadoum, M. (2011). Mechanisms Gating the Flow of Information in the Cortex: What They Might Look Like and What Their Uses may be. *Frontiers in Computational Neuroscience*, 5:1.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Gou, J., Yu, B., Maybank, S. J., and Tao, D. (2021). Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Güçlü, U. and Gerven, M. A. J. v. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, 35(27):10005–10014.

Hadsell, R., Rao, D., Rusu, A. A., and Pascanu, R. (2020). Embracing Change: Continual Learning in Deep Neural Networks. *Trends in Cognitive Sciences*, 24(12):1028–1040.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034.

Hebb, D. O. (1939). Intelligence in Man after Large Removals of Cerebral Tissue: Report of Four Left Frontal Lobe Cases. *The Journal of General Psychology*, 21(1):73–87.

Hebb, D. O. (1942). The effect of early and late brain injury upon test scores, and the nature of adult intelligence. *Proceedings of the American Philosophical Society*, 85:275–292.

Herd, S. A., O'Reilly, R. C., Hazy, T. E., Chatham, C. H., Brant, A. M., and Friedman, N. P. (2014). A neural network model of individual differences in task switching abilities. *Neuropsychologia*, 62:375–389.

Ito, T. and Murray, J. D. (2021). Multi-task representations in human cortex transform along a sensory-to-motor hierarchy. *bioRxiv*, page 2021.11.29.470432.

Iyer, A., Grewal, K., Velu, A., Souza, L. O., Forest, J., and Ahmad, S. (2022). Avoiding Catastrophe: Active Dendrites Enable Multi-Task Learning in Dynamic Environments. *Frontiers in Neurorobotics*, 16:846219.

Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. *arXiv:1806.07572 [cs, math, stat]*.

Jagadeesh, A. V. and Gardner, J. L. (2022). Texture-like representation of objects in human visual cortex. *bioRxiv*, page 2022.01.04.474849.

Jazayeri, M. and Ostojic, S. (2021). Interpreting neural computations by examining intrinsic and embedding dimensionality of neural activity. *Current Opinion in Neurobiology*, 70:113–120.

Johnston, K., Levin, H. M., Koval, M. J., and Everling, S. (2007). Top-down control-signal dynamics in anterior cingulate and prefrontal cortex neurons following task switching. *Neuron*, 53(3):453–462.

Kaplanis, C., Shanahan, M., and Clopath, C. (2018). Continual Reinforcement Learning with Complex Synapses. *arXiv:1802.07239 [cs]*.

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

Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, 10(11):e1003915.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Koay, S. A., Charles, A. S., Thiberge, S. Y., Brody, C. D., and Tank, D. W. (2022). Sequential and efficient neural-population coding of complex task information. *Neuron*, 110(2):328–349.e11.

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

Koechlin, E. and Summerfield, C. (2007). An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2.

Kurtz, K. H. and Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, 51(4):239–243.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *arXiv:1604.00289*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324. Conference Name: Proceedings of the IEEE.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2018). Deep Neural Networks as Gaussian Processes. *arXiv:1711.00165 [cs, stat]*.

Lee, J., Xiao, L., Schoenholz, S. S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. *arXiv:1902.06720 [cs, stat]*.

Lee, S., Goldt, S., and Saxe, A. (2021). Continual Learning in the Teacher-Student Setup: Impact of Task Similarity. In *Proceedings of the 38th International Conference on Machine Learning*, pages 6109–6119. PMLR.

Lhermitte, F. (1983). 'Utilization behaviour' and its relation to lesions of the frontal lobes. *Brain*, 106(2):237–255.

Libby, A. and Buschman, T. J. (2021). Rotational dynamics reduce interference between sensory and memory representations. *Nature Neuroscience*, pages 1–12.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7(1):13276.

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *Journal of Cognitive Neuroscience*, 33(10):2017–2031.

Liston, C., Matalon, S., Hare, T. A., Davidson, M. C., and Casey, B. J. (2006). Anterior Cingulate and Posterior Parietal Cortices Are Sensitive to Dissociable Forms of Conflict in a Task-Switching Paradigm. *Neuron*, 50(4):643–653.

Logothetis, N. K. (2003). The Underpinnings of the BOLD Functional Magnetic Resonance Imaging Signal. *Journal of Neuroscience*, 23(10):3963–3971.

Luria, A., Karpov, B., and Yarbuss, A. (1966). Disturbances of Active Visual Perception with Lesions of the Frontal Lobes. *Cortex*, 2(2):202–212.

Luria, A. R. (1973). Chapter 1 - The frontal lobes and the regulation of behaviour. In Pribram, K. H. and Luria, A. R., editors, *Psychophysiology of the Frontal Lobes*, pages 3–26. Academic Press.

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

Mansouri, F. A., Matsumoto, K., and Tanaka, K. (2006). Prefrontal Cell Activities Related to Monkeys' Success and Failure in Adapting to Rule Changes in a Wisconsin Card Sorting Test Analog. *Journal of Neuroscience*, 26(10):2745–2756.

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

Masse, N. Y., Grant, G. D., and Freedman, D. J. (2018). Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. *Proceedings of the National Academy of Sciences*, 115(44):E10467–E10475.

Matthews, P. M. and Jezzard, P. (2004). Functional magnetic resonance imaging. *Journal of Neurology, Neurosurgery & Psychiatry*, 75(1):6–12.

McClelland, J. L., McNaughton, B. L., and 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):419–457.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Bower, G. H., editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Miller, E. K. (2000). The prefontral cortex and cognitive control. *Nature Reviews Neuroscience*, 1(1):59–65.

Miller, E. K. and Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24:167–202.

Miller, E. K., Freedman, D. J., and Wallis, J. D. (2002). The prefrontal cortex: categories, concepts and cognition. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 357(1424):1123–1136.

Miller, G. A., Galanter, E., and Pribram, K. H. (1960). *Plans and the structure of behavior*. Plans and the structure of behavior. Henry Holt and Co, New York, NY, US.

Minsky, M. (1961). Steps toward Artificial Intelligence. *Proceedings of the IRE*, 49(1):8–30.

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

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

Musslick, S. and Cohen, J. D. (2021). Rationalizing constraints on the capacity for cognitive control. *Trends in Cognitive Sciences*, 25(9):757–775.

Musslick, S., Saxe, A., Hoskin, A. N., Reichman, D., and Cohen, J. D. (2020). On the Rational Boundedness of Cognitive Control: Shared Versus Separated Representations. *PsyArXiv*.

Musslick, S., Saxe, A., Özcimder, K., Dey, B., Henselman, G., and Cohen, J. D. (2017). Multitasking capability versus learning efficiency in neural network architectures. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society (pp. 829–834)*. ISSN: 1069-7977.

Nelli, S., Braun, L., Dumbalska, T., Saxe, A., and Summerfield, C. (2021). Neural knowledge assembly in humans and deep networks. *bioRxiv*, page 2021.10.21.465374.

Newell, A., Shaw, J. C., and Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3):151–166.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLOS Computational Biology*, 10(4):e1003553.

Norman, D. A. and Shallice, T. (1986). Attention to Action. In Davidson, R. J., Schwartz, G. E., and Shapiro, D., editors, *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4*, pages 1–18. Springer US, Boston, MA.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273.

Oja, E. and Karhunen, J. (1985). On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84.

O'Reilly, R. C. and Frank, M. J. (2006). Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328.

Paccolat, J., Petrini, L., Geiger, M., Tyloo, K., and Wyart, M. (2021). Geometric compression of invariant manifolds in neural nets. *arXiv:2007.11471 [cs, stat]*.

Panichello, M. F. and Buschman, T. J. (2021). Shared mechanisms underlie the control of working memory and attention. *Nature*, 592(7855):601–605.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Pavlov, I. P. (1927). *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford Univ. Press, Oxford, England.

Poldrack, R. A. (2008). The role of fMRI in Cognitive Neuroscience: where do we stand? *Current Opinion in Neurobiology*, 18(2):223–227.

Posner, M. and Snyder, C. (1975). Attention and cognitive control. In *Information processing and cognition: The loyola symposium*. Erlbaum, Hillsdale NJ.

Posner, M. I. and Presti, D. E. (1987). Selective attention and cognitive control. *Trends in Neurosciences*, 10(1):13–17.

Rainer, G., Asaad, W. F., and Miller, E. K. (1998). Selective representation of relevant information by neurons in the primate prefrontal cortex. *Nature*, 393(6685):577–579.

Raposo, D., Kaufman, M. T., and Churchland, A. K. (2014). A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17(12):1784–1792.

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., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A. C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., and Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770.

Ridley, R. M. (1994). The psychology of perseverative and stereotyped behaviour. *Progress in Neurobiology*, 44(2):221–231.

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

Rikhye, R. V., Gilra, A., and Halassa, M. M. (2018). Thalamic regulation of switching between cortical representations enables cognitive flexibility. *Nature Neuroscience*, 21(12):1753–1763.

Rizzolatti, G., Riggio, L., Dascola, I., and Umiltá, C. (1987). Reorienting attention

across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia*, 25(1, Part 1):31–40.

Robinson, D. A. and Fuchs, A. F. (1969). Eye movements evoked by stimulation of frontal eye fields. *Journal of Neurophysiology*, 32(5):637–648.

Rohrer, D., Dedrick, R. F., and Stershic, S. (2015). Interleaved practice improves mathematics learning. *Journal of Educational Psychology*, 107(3):900–908.

Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T. P., and Wayne, G. (2019). Experience Replay for Continual Learning. *arXiv:1811.11682 [cs, stat]*.

Rorden, C. and Karnath, H.-O. (2004). Using human brain lesions to infer function: a relic from a past era in the fMRI age? *Nature Reviews Neuroscience*, 5(10):812–819.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408.

Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D., and O'Reilly, R. C. (2005). Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102(20):7338–7343.

Roy, J. E., Riesenhuber, M., Poggio, T., and Miller, E. K. (2010). Prefrontal Cortex Activity during Flexible Categorization. *Journal of Neuroscience*, 30(25):8519–8528.

Rule, M. E., O'Leary, T., and Harvey, C. D. (2019). Causes and consequences of representational drift. *Current Opinion in Neurobiology*, 58:141–147.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Rumelhart, D. E. and McClelland, J. L. (1987). A General Framework for Parallel Distributed Processing. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*, pages 45–76. MIT Press.

Russin, J., Zolfaghar, M., Park, S. A., Boorman, E., and O'Reilly, R. C. (2022). A Neural Network Model of Continual Learning with Cognitive Control. *arXiv:2202.04773 [cs, q-bio]*.

Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. (2016). Progressive Neural Networks. *arXiv:1606.04671 [cs]*.

Sadtler, P. T., Quick, K. M., Golub, M. D., Chase, S. M., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M., and Batista, A. P. (2014). Neural constraints on learning. *Nature*, 512(7515):423–426. Number: 7515 Publisher: Nature Publishing Group.

Samani, J. and Pan, S. C. (2021). Interleaved practice enhances memory and problem-solving ability in undergraduate physics. *npj Science of Learning*, 6(1):1–11.

Sato, T. R. and Schall, J. D. (2003). Effects of Stimulus-Response Compatibility on Neural Selection in Frontal Eye Field. *Neuron*, 38(4):637–648.

Saxe, A. (2015). *Deep Linear Networks: A Theory of Learning in the Brain and Mind.* PhD thesis, Stanford University.

Saxe, A., Nelli, S., and Summerfield, C. (2021). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546.

Saxena, S. and Cunningham, J. P. (2019). Towards the neural population doctrine. *Current Opinion in Neurobiology*, 55:103–111.

Schafer, R. J. and Moore, T. (2007). Attention Governs Action in the Primate Frontal Eye Field. *Neuron*, 56(3):541–551.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., and

DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, page 407007.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038.

Sheahan, H., Luyckx, F., Nelli, S., Teupe, C., and Summerfield, C. (2021). Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, 109(7):1214–1226.e8.

Shiffrin, R. and Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, 84(2):127–190.

Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual Learning with Deep Generative Replay. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Siegel, M., Buschman, T. J., and Miller, E. K. (2015). Cortical information flow during flexible sensorimotor decisions. *Science*, 348(6241):1352–1355.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, 57(4):193–216.

Soetens, E., Boer, L. C., and Hueting, J. E. (1985). Expectancy or automatic facilitation? Separating sequential effects in two-choice reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, 11(5):598–616.

Sohl-Dickstein, J., Novak, R., Schoenholz, S. S., and Lee, J. (2020). On the infinite width limit of neural networks with a standard parameterization. *arXiv:2001.07301 [cs, stat]*.

Storrs, K. R. and Kriegeskorte, N. (2019). Deep Learning for Cognitive Neuroscience. *arXiv*.

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

Summerfield, C., Luyckx, F., and Sheahan, H. (2020). Structure learning and the posterior parietal cortex. *Progress in Neurobiology*, 184:101717.

Takagi, Y., Hunt, L. T., Woolrich, M. W., Behrens, T. E., and Klein-Flügge, M. C. (2021). Adapting non-invasive human recordings along multiple task-axes shows unfolding of spontaneous and over-trained choice. *eLife*, 10:e60988.

Tang, E., Mattar, M. G., Giusti, C., Lydon-Staley, D. M., Thompson-Schill, S. L., and Bassett, D. S. (2019). Effective learning is accompanied by high-dimensional and efficient representations of neural activity. *Nature Neuroscience*, 22(6):1000–1009.

Thagard, P. (2020). Cognitive Science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2020 edition.

Thorndike, E. L. (1932). *The fundamentals of learning*. The fundamentals of learning. Teachers College Bureau of Publications, New York, NY, US.

Tosoni, A., Galati, G., Romani, G. L., and Corbetta, M. (2008). Sensory-motor mechanisms in human parietal cortex underlie arbitrary visual decisions. *Nature neuroscience*, 11(12):10.1038/nn.2221.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136.

Tsuda, B., Tye, K. M., Siegelmann, H. T., and Sejnowski, T. J. (2020). A modeling framework for adaptive lifelong learning with transfer and savings through gating in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, 117(47):29872–29882.

van de Ven, G. M., Siegelmann, H. T., and Tolias, A. S. (2020). Brain-inspired replay for continual learning with artificial neural networks. *Nature Communications*, 11(1):4069.

van de Ven, G. M. and Tolias, A. S. (2019). Three scenarios for continual learning. *arXiv:1904.07734 [cs, stat]*.

Verbeke, P. and Verguts, T. (2022). Using top-down modulation to optimally balance shared versus separated task representations. *Neural Networks*, 146:256–271.

Voulodimos, A., Doulamis, N., Doulamis, A., and Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018:e7068349.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review*, 20(2):158–177.

Werbos, P. (1974). *Beyond Regression : "New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD Dissertation, Harvard University.

Woodward, T. S., Ruff, C. C., and Ngan, E. T. C. (2006). Short- and long-term changes in anterior cingulate activation during resolution of task-set competition. *Brain Research*, 1068(1):161–169.

Woodworth, B., Gunasekar, S., Savarese, P., Moroshko, E., Golan, I., Lee, J., Soudry, D., and Srebro, N. (2020). Kernel and Rich Regimes in Overparametrized Models. *arXiv:1906.05827 [cs, stat]*.

Wulf, G. and Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, 9(2):185–211.

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., and Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, 375(6581):632–639.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624.

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

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

Yu, A. J. and Cohen, J. D. (2008). Sequential effects: Superstition or rational behavior? In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience*, 16(8):487–497.

Zeng, G., Chen, Y., Cui, B., and Yu, S. (2019). Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372.

Zenke, F., Poole, B., and Ganguli, S. (2017). Continual Learning Through Synaptic Intelligence. *arXiv:1703.04200 [cs, q-bio, stat]*.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., and Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, 118(3).

Zohary, E., Shadlen, M. N., and Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143.

Appendix A

Supplementary Information for Chapter 3

A.1 Supplementary Methods

Impact of learning rate on representation learning. We tested whether changing the learning rate could push the network into either the rich or lazy regime. Starting from a lazy (weight scale 3) or rich (weight scale 0.01) initialisation, we trained the network with 20 independent runs per learning rate, which ranged from 0.001 to 0.01 in 10 steps. The values were chosen to ensure that learning dynamics remained stable.

Controlling the learning regime via L2-regularisation. We investigated whether a network initialised in the lazy regime could be pushed into the rich regime by adding a regularisation term that favoured small weights. For this, we added an L2 regulariser to the loss function:

$$J(W) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i, W))^2 + \lambda ||W||_2^2 \tag{A.1}$$

The hyperparamter λ controlled the regularisation strength. We initialised the network in the lazy regime (weight scale $\sigma = 3$) and collected 30 independent runs per regularisation strength, which ranged from 0 to 0.1.

A.2 Supplementary Figures

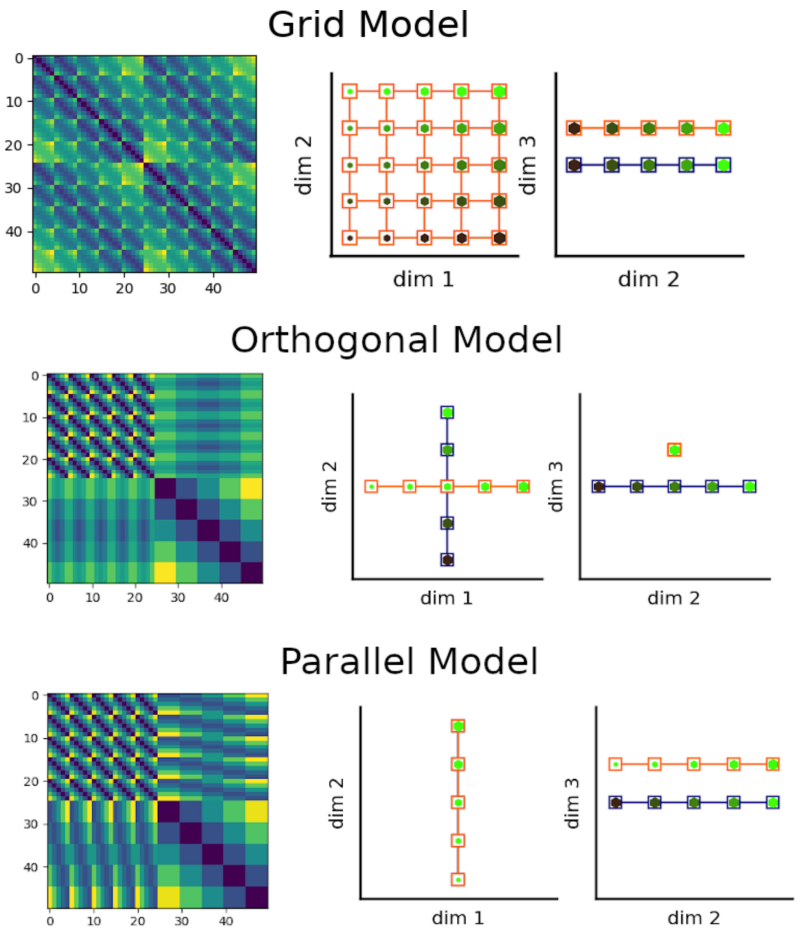

Figure A.1: Model RDMs. The grid model encoded both feature dimensions and the context along separate axes, forming a 5x5x2 grid of evenly spaced stimuli. The orthogonal model was performed by projecting the grid model onto the task-relevant axes, leaving a representation in which the stimuli fall onto two orthogonal lines, one for each task. A separate dimension encodes the context. The parallel model was obtained by rotating one of the lines from the orthogonal model by 90 degrees, so that stimuli were aligned according to their signed distance to the context-specific category bound.

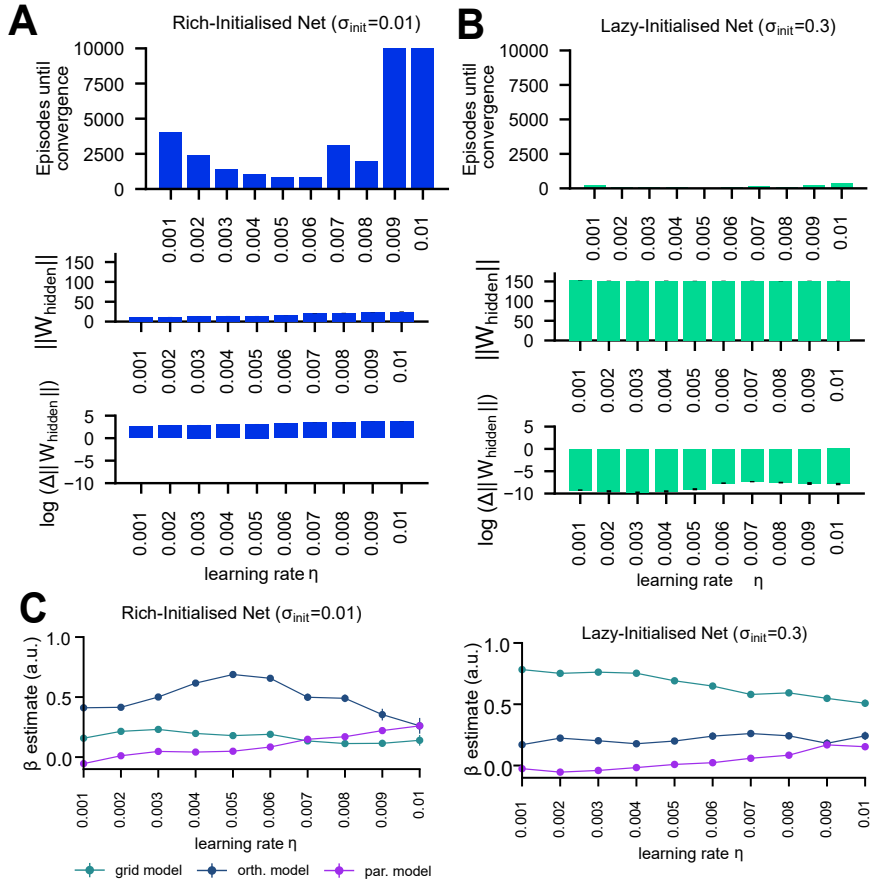

Figure A.2: Impact of different learning rates on representations. (**A**) Time until convergence (top) and weight change (middle/bottom) as a function of the learning rate for a network initialized in the rich regime. (**B**) Same as (A) but initialized in lazy regime. Note that irrespective of learning rate, all nets converged faster than under (A) and had smaller weight changes. (**C**) RSA on hidden layer patterns as function of learning rate, for rich (top) and lazy (bottom) initialized nets, showing again that learning rate was not critical.

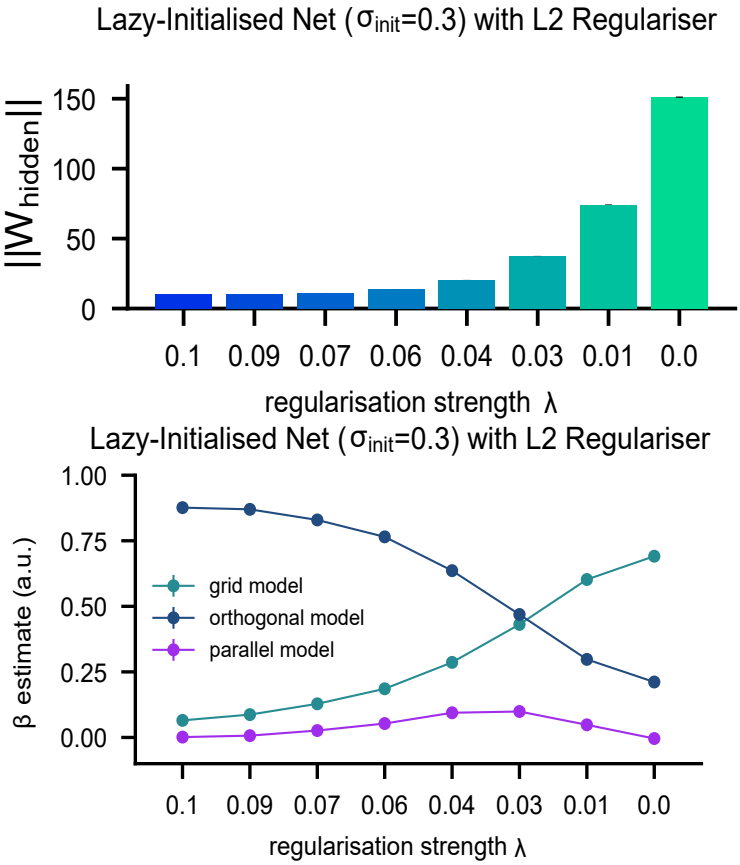

Figure A.3: Using L2 regularisation to induce rich learning. Lazy-initialised network with L2 regulariser, demonstrating that weight norm (top) and task-specificity of representations (bottom) can be controlled by changing the regularisation strength.

Appendix B

Supplementary Information for Chapter 4

B.1 Supplementary Methods

Recurrent Neural Network Extension: Let $x_1(t) \in [-1,1]$ be the signed motion coherence over time in a trial, and $x_2(t) \in [-1,1]$ be the signed color level over time, which can be stacked into the column vector input $x(t) = [x_1(t)\ x_2(t)]^T$. Let $u(t) \in \mathbb{R}^2$ be the task context input encoded as a one hot vector (+1 in the first element for context A, +1 in the second element for context B).

The network contains four neuron classes, and the overall architecture is depicted in **Fig. B.7**. In particular, these comprise a pair selective for positive/negative motion and task, and an pair selective for positive/negative color and task. Each neuron receives stimulus input through the input-to-hidden weights:

$$W_x = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \tag{B.1}$$

Each neuron class also receives task input, with the motion neurons receiving inhibitory input in the color task and the color neurons receiving inhibitory input in

the motion task. The task-to-hidden weights are

$$W_u = \begin{bmatrix} 0 & -w \\ 0 & -w \\ -w & 0 \\ -w & 0 \end{bmatrix} \tag{B.2}$$

where w is a parameter controlling the strength of context-driven inhibition.

The network has recurrence, which we assume has an autapse structure such that each neuron has self recurrence with weight one to enable persistent activity.

We emphasize that all four neuron classes are mixed selective, in the sense that their response depends on a combination of stimulus and task. However, this mixed selectivity is not random, rather it is highly structured.

The neural activity dynamics are given by the standard firing rate equations

$$\frac{d}{dt}h(t) = -h(t) + f(h(t) + W_x x(t) + W_u u(t)) \tag{B.3}$$

where $f(\cdot)$ is the firing rate nonlinearity, which here we take to be the ReLU function ($f(v) = \max\{v, 0\}$).

Finally the output of the network r is computed through readout weights $W_o = [1 \ -1 \ 1 \ -1]$, i.e., by summing or subtracting the relevant hidden unit activity,

$$r(t) = W_o h(t) \tag{B.4}$$

We now describe the temporal structure of a trial. We assume that between trials, neural activity resets such that we have the initial condition $h(0) = 0$. We assume that input stimuli arrive with a temporal profile $p_x(t)$ that is rescaled by the motion coherence m and color coherence c, such that the input is

$$x(t) = \begin{bmatrix} m p_x(t) \\ c p_x(t) \end{bmatrix} \tag{B.5}$$

For simplicity we take $p_x(t) = ae^{-t/\tau} + b$ for $0 < t < t_x$, and $p_x(t) = 0$ otherwise, to reflect a sharp onset transient followed by decay to a steady state.

The context signal arrives with a temporal profile $p_u(t)$, turning on with the stimulus and remaining on during the delay period until some time $t_u > t_x$. For simplicity we take $p_u(t)$ to be a pulse (one for times between 0 and t_u, zero otherwise). Let z be 1 in the motion context and 0 in the color context. Then we have

$$u(t) = \begin{bmatrix} z p_u(t) \\ (1-z)\, p_u(t) \end{bmatrix} \tag{B.6}$$

B.2 Supplementary Figures

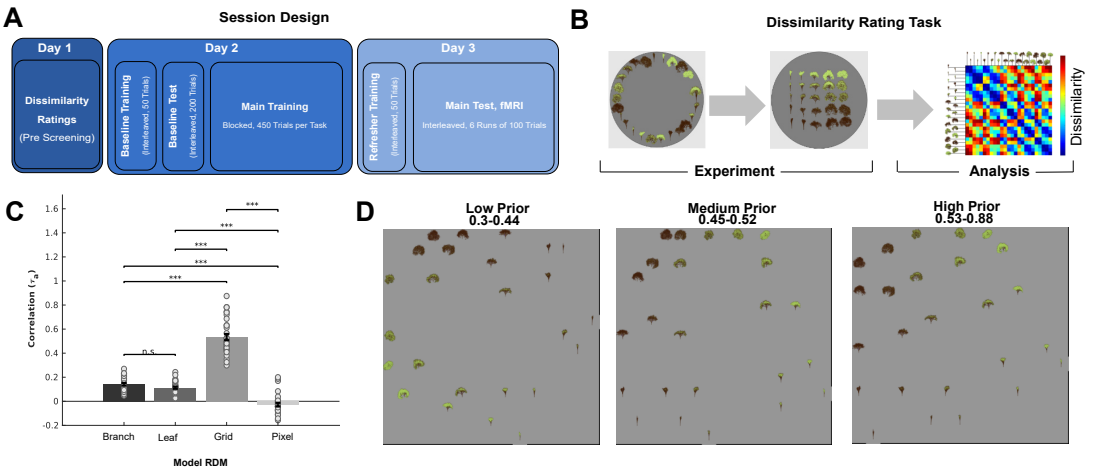

Figure B.1: Experimental design and grid prior analysis. (**A**) Session Design. Participants completed three sessions carried out over consecutive days. All participants underwent a screening task (day1) in which they were asked to perform dissimilarity ratings on tree stimuli. Those who showed strong evidence for being aware of the dimensions of branchiness and leafiness (assessed by a "grid score", see next figure) were invited to the remaining parts of the study. On day 2, participants received a lengthy blocked training curriculum, preceded by a brief familiarisation phase and evaluation (baseline training and test) to measure the effectiveness of the training phase. On day 3, participants received a brief refresher training, before they underwent fMRI scanning during which they completed six interleaved blocks of test trials. See methods sections for additional details. (**B**) Dissimilarity Rating Task & RSA. Participants were asked to arrange tree stimuli via mouse drag & drop in a circular arena such that distances between trees corresponded to how dissimilar they were perceived (left and middle panel). From these ratings, we constructed RDMs at single subject level. These RDMs were correlated with model RDMs assuming that participants were (i) only aware of branchiness, (ii) only aware of leafiness, (iii) aware of the full 5x5 grid of branchiness and leafiness or (iv) made judgements based on pixel similarity. We describe the extent to which the third model explains the data as "grid score". In Flesch et al, 2018, we reported interactions between training effectiveness and grid score. We thus only invited participants with a grid score higher than the median grid score (tau=0.18) from the previous study. All screened participants exceeded this threshold. (**C**) Correlation coefficients between subject ratings and model RDMs. The grid model explained the data best, indicating that participants were on average aware of the data-generating dimensions. (**D**) MDS on dissimilarity ratings, divided into participants with low, medium and high grid score. All groups showed evidence for awareness of the dimensions branchiness & leafiness, and their grid-like relationship with each other.

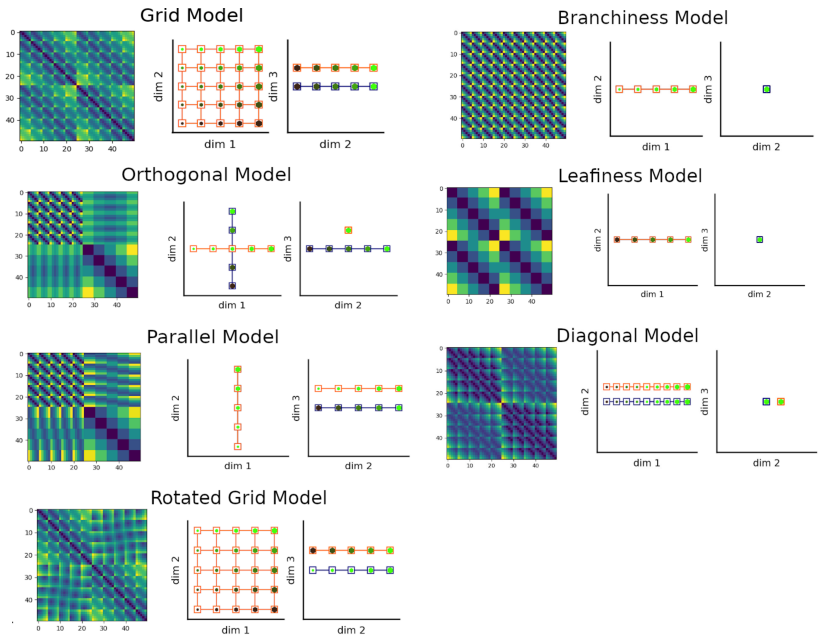

Figure B.2: Model RDMs. The grid model encoded both feature dimensions and the context along separate axes, forming a 5x5x2 grid of evenly spaced stimuli. The orthogonal model was performed by projecting the grid model onto the task-relevant axes, leaving a representation in which the stimuli fall onto two orthogonal lines, one for each task. A separate dimension encodes the context. The parallel model was obtained by rotating one of the lines from the orthogonal model by 90 degrees, so that stimuli were aligned according to their signed distance to the context-specific category bound. The branchiness and leafiness models encode only a single feature, irrespective of the context. The diagonal model was obtained by projecting stimuli onto the main diagonal running from low branchiness and leafiness to high branchiness and leafiness. The rotated grid model is similar to the parallel model, in the sense that one of the grids from the original grid model were rotated by 90 degrees so that both task representations are in the frame of reference of the response.

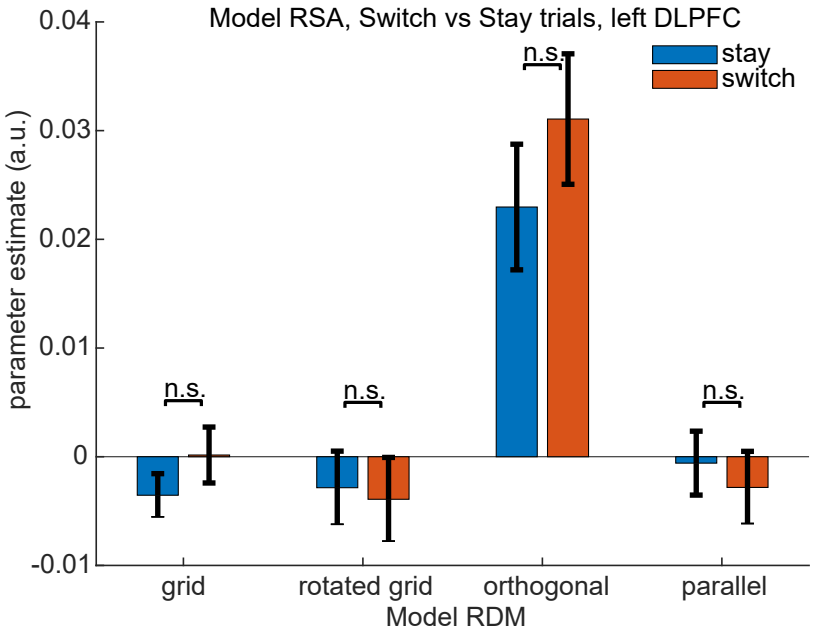

Figure B.3: Switch cost and task factorisation The univariate contrast of switch vs stay trials revealed a significant difference in BOLD in left DLPFC, an area where we had also observed evidence for factorised representations using the searchlight RSA approach. We therefore tested whether the extent to which task representations were factorised (orthogonal manifolds) differed between switch and stay trials. The difference, however, was not significant.

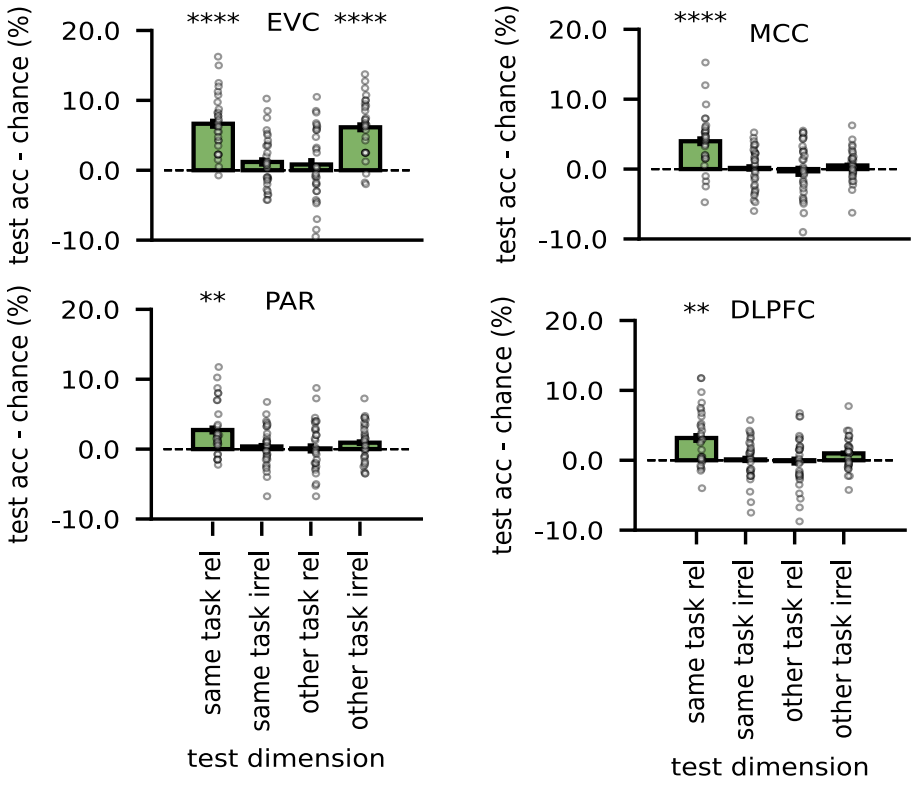

Figure B.4: Decoding of feature dimensions from fMRI data Cross-validated decoding accuracies for a linear SVM trained on the relevant dimension on one task and evaluated on the relevant/irrelevant dimension of the same and the other task. Only in EVC, the same dimension can be decoded from the other task (where it was irrelevant), showing that the other regions suppressed task-irrelevant dimensions. Asterisks indicate p-values after Bonferroni-correction.

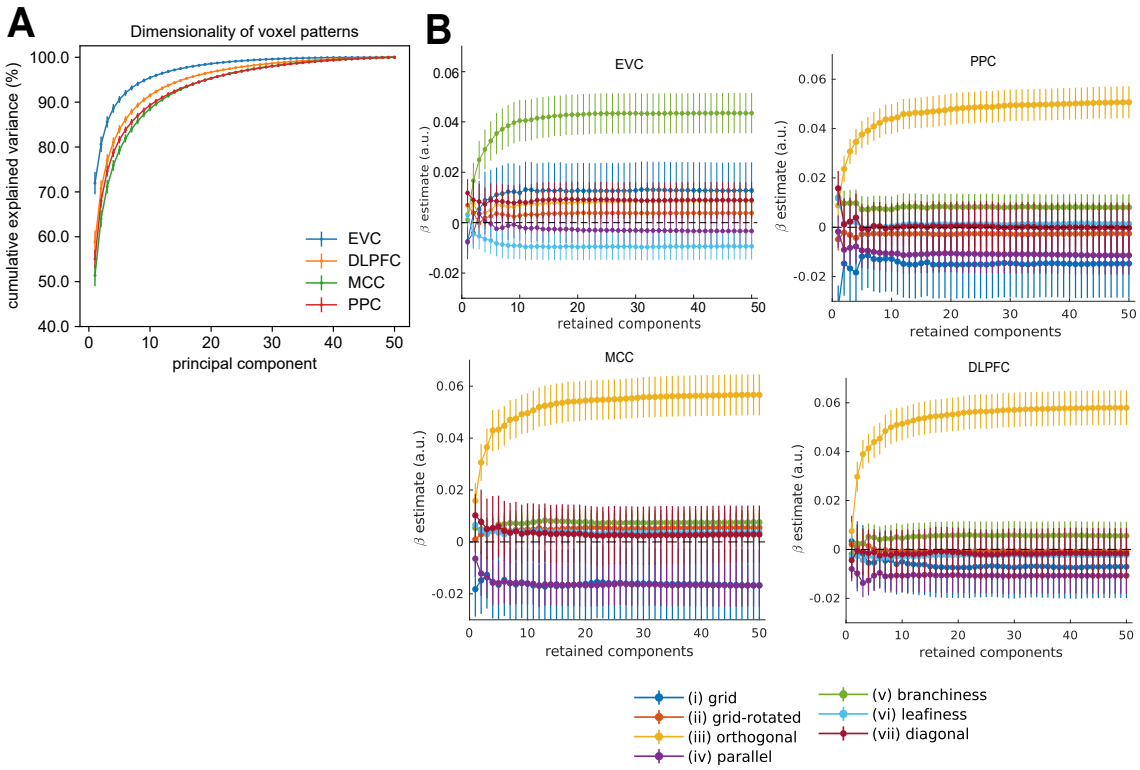

Figure B.5: Dimensionality of neural representations (**A**) Pattern dimensionality in the four ROIs. Scree plot of the cumulative explained variance in the four ROIs. (**B**) Truncated SVD. We repeated the RSA after successively removing principal components from the activity patterns in each ROI, establishing a lower bound on the number of components required for successfully decoding grid-like or orthogonal patterns in EVC and fronto-parietal regions respectively.

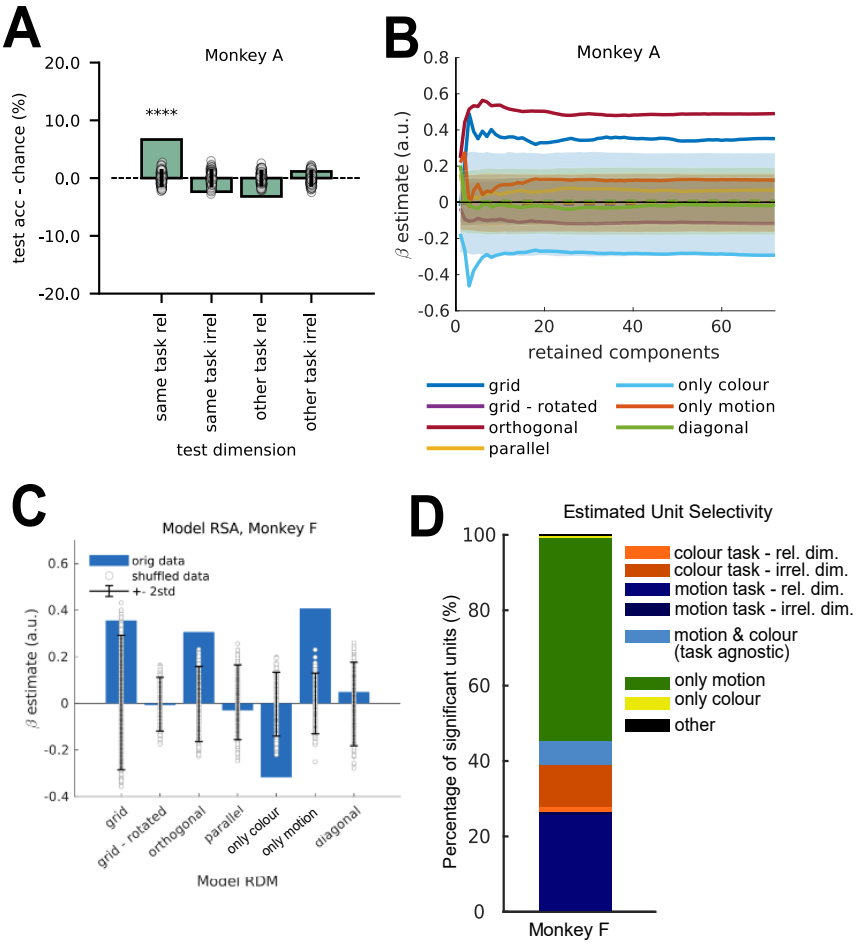

Figure B.6: Supplementary analyses of NHP data. (**A**) Cross-validated decoding of feature dimensions on pseudo-trials generated from Monkey A data (methods), showing that only task-relevant dimensions were represented in the neural pattern. (**B**) Truncated SVD on patterns from monkey A, showing that the orthogonal model explains the data even if only the top 3 principal components are maintained. (**C**) RSA for Monkey F, showing that in contrast to monkey A (main text), patterns encoded predominantly motion, irrespective of context. (**D**) Distribution of unit selectivity for monkey F; most recorded units are motion-selective.

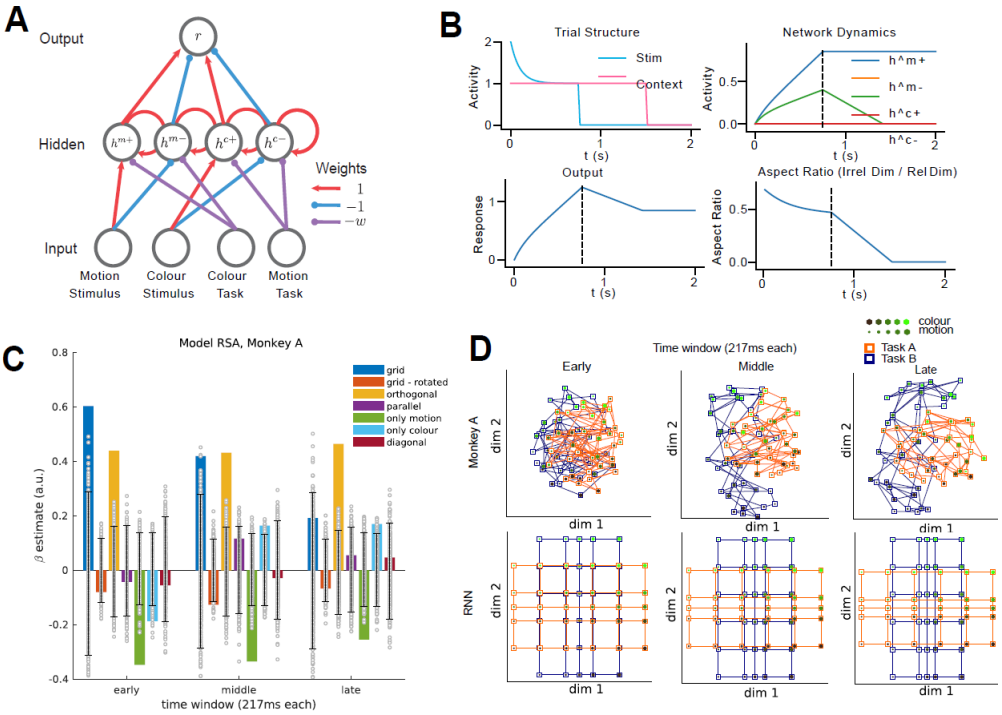

Figure B.7: RNN model extension. (**A**) RNN version of our neural network model. (**B**) RNN dynamics throughout a simulated trial. Top left: Stimulus and context are presented for 750ms, followed by delay (1s) where only context is present. Top right: Hidden layer dynamics during motion task trial. We observe a gradual integration of motion information in the motion-sensitive unit, and, to a lesser extent, colour information in the colour-sensitive unit. After stimulus offset (dashed line), the irrelevant dimension (colour) is gradually suppressed by the context signal. Bottom left: Gradual integration of a category signal in the output unit, which remains roughly constant after stimulus offset. Bottom right: Aspect ratio between activity encoding the irrelevant and relevant dimensions respectively, indexing the amount of compression along irrelevant dimensions. The aspect ratio decreases during the stimulus interval as irrelevant and relevant feature information are integrated at different rates (top right plot). It decreases more rapidly after stimulus offset (dashed line) as the context signal filters out any task-irrelevant information that is still present. (**C**) Model RSA on NHP data, separately for early, middle and late time windows within the stimulus interval, suggesting that the neural code transforms from a grid-like to an orthogonal and task-specific representation. (**D**) MDS on monkey and RNN RDMs averaged over early, middle and late time windows within the stimulus interval.

Table T1. Results of Bayesian Model Comparison between brain regions

ROI		**Protected Exceedance Probability**			**Estimated Model Frequencies**	
	Branchiness & Orthogonal	Branchiness	Orthogonal	Branchiness & Orthogonal	Branchiness	Orthogonal
EVC	0.302	**0.489**	0.209	0.38±0.08, z=0.8, p=0.21	**0.5±0.08 z=2.29, p=0.01**	0.12±0.05 z=-3.08, p=0.999
DLPFC	0.0	0.0	**1.0**	0.15±0.06, z=-2.86, p=0.99	0.03±0.02, z=-4.28, p=1	**0.82±0.06, z=4.51, p<0.0001**
MCC	0.003	0.002	**0.995**	0.20±0.06, z=-2.00, p=0.97	0.07±0.04, z=-3.69, p=1	**0.73±0.07, z=4.19, p<0.0001**
PPC	0.0	0.0	**1.0**	0.12±0.05, z=-3.08, p=0.99	0.04±0.03, z=-4.24, p=1	**0.84±0.06, z=4.57, p<0.0001**

Figure B.8: Results of Bayesian Model Comparison between brain regions

Appendix C

Supplementary Material for Chapter 6

C.1 Supplementary Methods

Stimulus norming task

To assess the validity of the size and speed levels we had assigned to each animal/vehicle of the stimulus dataset, we conducted a norming study in which we asked separate groups of participants to rate the size/speed of animals/vehicles on a five-point Likert scale.

Participants: We recruited a total of 120 participants (30 per group, 24 male, 72 female, 1 other, mean age 27.14 years) on the crowdsourcing platform Prolific. Participants were compensated for their time at a rate of £10/hour. We restricted recruitment to participants in the age range 18-40 who were ordinarily residents in the UK and had an approval rating of at least 85% averaged over their past five submissions. All experiments were approved by the Medical Sciences Research Ethics Committee of the University of Oxford (approval reference: R50750/RE001). No participants were excluded from this experiment.

Stimulus Design: We collected images of 25 different animal types and 25 different vehicle types from public databases and via Google image search, so that stimuli within each domain would span a 5x5 grid of different speed and size levels. For each type, we collected at least 10 unique exemplars. Images were pre-screened to ensure that physical size in terms of proportion of pixels occupied by the animal/vehicle were approximately similar across all levels. All images were cropped and rescaled to

400x200 pixels.

Task Design: Participants were asked to rate stimuli according to their size or speed on a five-step Likert scale. We tested four groups in total, one per combination of domain (animals/vehicles) and feature (size/speed). On each trial, participants saw an image of an animal/vehicle in the centre of the screen, together with a Likert scale displayed just below the image. The items on the scale were labelled "very slow, slow, medium, fast, very fast" for the speed task and "very small, small, medium, large, very large" for the size task. The experiment was self-paced, but participants could only proceed with the next trial once they had submitted a response. Each participant completed 250 trials, with 10 trials per combination of size and speed and trial-unique stimuli (for example, 10 different images of a leopard).

Quantification and statistical analysis: Ratings were averaged within participants by taking the median across the 10 trial-unique stimuli per combination of size and speed. We then plotted the "ground truth" labels against the group-level average of these subjective ratings, and computed the rank-correlation (Kendall's tau a) to quantify the agreement between the labels and the ratings given by the participants.

Neural network simulations

To study the geometry of task representations, we trained neural networks on interleaved trials from a synthetic dataset with two domains and two categorisation rules per domain, similar to the animals/vehicles study. In contrast to previous simulations, inputs were trial-unique random vectors, so that all information about size/speed was contained in the labels. We predicted that different rules, such as speed versus size, would be mapped onto orthogonal axes in neural state space, both within and across domains, while representations of the same rules, such as speed in the first domain versus speed in the second domain, should lie on parallel planes to facilitate cross-domain generalisation.

Network Architecture: The network was a simple feed-forward MLP with two hidden layers (1000 units each, ReLU-nonlinearities) and linear output.

Task Design: The dataset consisted of 100 random binary vectors (25 per task and domain). As in previous simulations, we provided the network with a contextual cue,

which consisted of a four-dimensional vector, coding for the four tasks (size/speed of animals/vehicles). Labels were provided by calculating the cartesian product between five levels of "speed" and "size" and assigning either the size or speed level (whichever was the relevant dimension) to the 25 binary vectors in each task. Hence, in contrast to previous simulations, information about the size/speed of the stimuli was not encoded in the stimuli but provided solely by the labels.

Training Procedure: We trained the network with online SGD on the training data described above, using an MSE loss on the difference between the ground-truth labels (size or speed) and the network's outputs. The network was initialised with small weights, bringing it in the rich training regime (weight variance $\sigma^2 = \frac{1}{sqrt(n_{hidden})}$). The learning rate was set to $\varepsilon = 1e-2$ for all simulations and training was carried out for 5000 epochs, where one epoch corresponds to a training run on all 100 unique training trials. For statistical inference, we collected 10 unique runs with random initialisations per simulation.

Quantification and statistical analysis: Analyses of the accuracy and hidden layer geometries were like those presented in previous chapters. For the RSA, however, we now constructed RDMs based on activity patterns form both domains, allowing us to investigate representational geometries within and across domains. Again, we tested for the presence of grid-like, orthogonal, or parallel representational geometries, by regressing the RDMs computed from hidden-layer activity post training against a set of model RDMs corresponding to these geometries.

C.2 Supplementary Figures

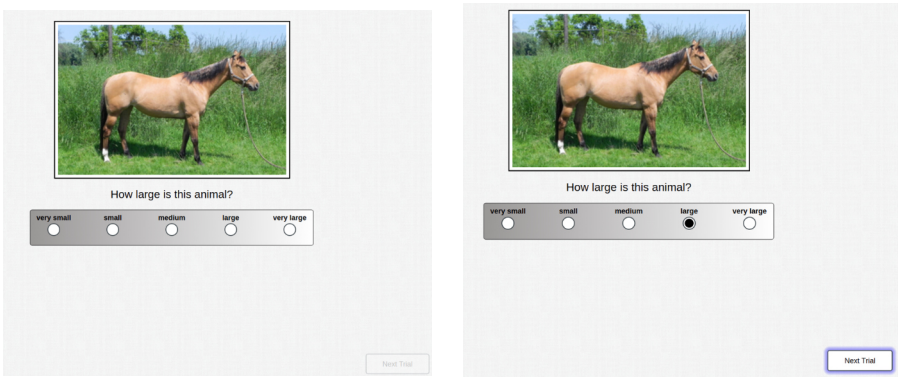

Figure C.1: Rating task (Stimulus norming study). On each trial, participants judged the speed/size of an animal/vehicle on a five point Likert scale.

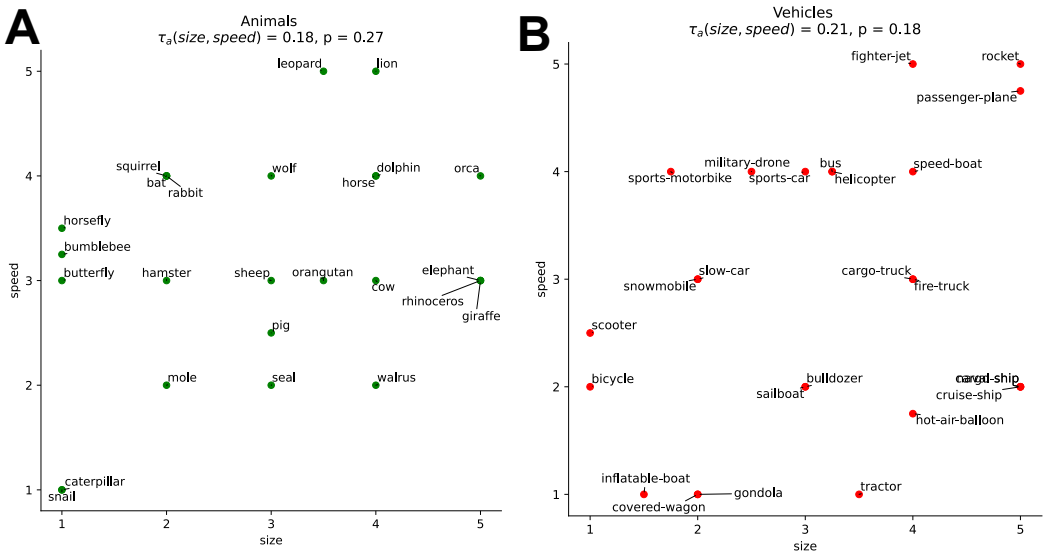

Figure C.2: Rating Task, Correlation between ratings(**A**) Correlation between ratings for size and speed, animals task. We constructed the stimulus space so that these two dimensions should be uncorrelated. Indeed, correlations between the two dimensions were only modest ($\tau_a(size, speed) = 0.18$). (**B**) Same as (A), but for vehicles, ($\tau_a(size, speed) = 0.21$).

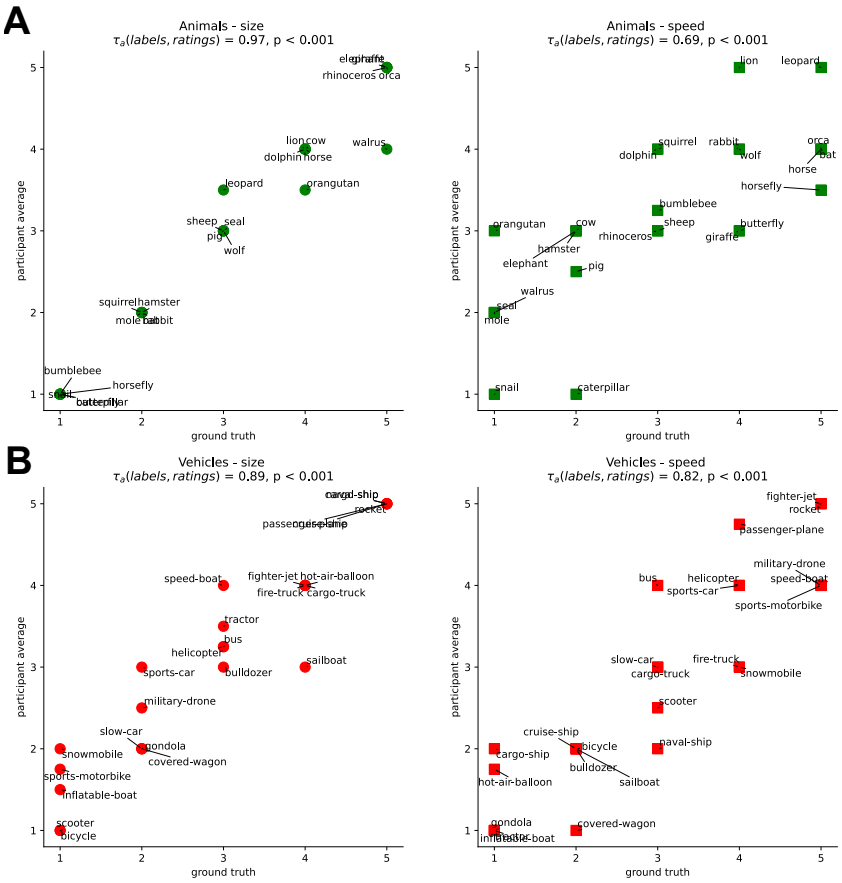

Figure C.3: Rating Task, correlation with ground truth. (**A-B**) Correlation between ratings and ground truth. Across domains and feature dimensions, we found evidence for strong agreement between the ground-truth labels we had assigned to the stimuli and the subjective ratings given by our participants (Animals-speed: 0.69; animals-size: 0.97; vehicles-speed: 0.89; vehicles-size: 0.82).

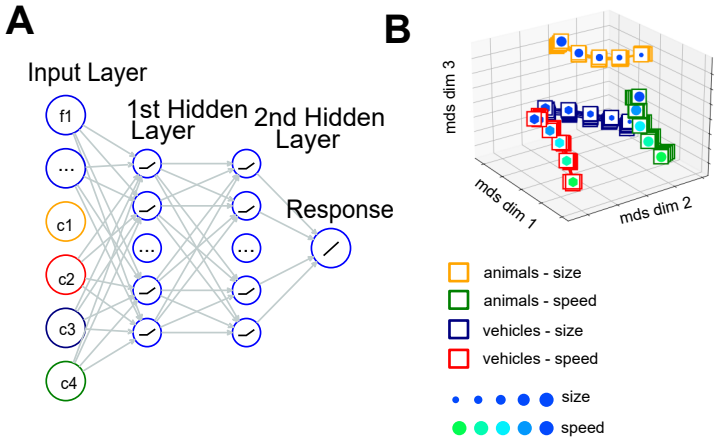

Figure C.4: Simulation of the two-domain task with interleaved data. (**A**) Network architecture. We trained a feedforward MLP with two hidden layers on the animals/vehicles tasks. Inputs were random binary vectors that represented stimuli, together with four nodes that coded for the relevant task and domain in a one-hot fashion. (**B**) Projection of the geometries in the second hidden layer into three dimensions, revealing that similar rules from different domains fall onto parallel coding axes, and that competing rules (size versus speed) are orthogonal to each other.