Sequential learning in the form of shaping as a source of cognitive flexibility

Kai A. Krueger

Gatsby Computational Neuroscience Unit
University College London
17 Queen Square
WC1N 3AR
London, United Kingdom

THESIS

Submitted for the degree of
Doctor of Philosophy, University of London

2011

I, Kai Arne Krueger, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Humans and animals have the ability to quickly learn new tasks, a rapidity that is unlikely to be manageable by pure trial and error learning on each task separately. Instead, key to this rapid adaptability appears to be the ability to integrate skills and knowledge obtained from previous tasks. This is assumed for example in the sequential build-up of curricula in education, and has been employed in training animals for behavioural experiments at least since the initial work on shaping by Skinner in 1938.

Despite its importance to natural learning, from a computational neuroscience point of view the question of sequential learning of tasks has largely been ignored. Instead, learning algorithms have often been devised that are capable of learning from an initial naïve state. However, it is known that simply training sequentially with the same algorithms can often *harm* learning through *interference*, rather than enhance it.

In this thesis, we explore the effects of sequential training in the form of shaping in the cognitive domain. We consider abstract, yet neurally inspired, learning models and propose extensions and requirements to ensure that shaping is beneficial.

We take the 12-AX task, a hierarchical working memory task with rich structure, define a shaping sequence to break the hierarchical structure of the task into separate smaller and simpler tasks, and compare performance between learning the task in one fell swoop to that of learning it with the help of shaping.

Using a Long Short-Term Memory (LSTM) network model, we show that learning times can be reduced substantially through shaping. Furthermore, other metrics such as forms of abstraction and generalisation may also show differential effects. Crucial to this, though, is the ability to prevent interference, which we achieve through an architectural extension in the form of "resource allocation".

Finally, we present initial, human behavioural data on the 12-AX task, showing that humans can learn it in a single session. Nevertheless, the task is sufficiently challenging to reveal interesting behavioural structure. This supports its use as a candidate to probe computational aspects of cognitive learning, including shaping. Furthermore, our data show that the shaping protocol used in the modelling studies can also improve averaged asymptotic performance in humans.

Overall, we show the importance of taking sequential task learning into account, provided there is additional architectural support. We propose and demonstrate candidates for this support.

Acknowledgements

I have, no doubt, many people to thank, as for without their help and support this thesis would not have been possible. There have been ups and downs, but always were there people for support, to share, to enjoy and to work with, making the experience at Gatsby overall a very pleasant and enjoyable environment.

Foremost, I would like to thank my supervisor Peter Dayan, without whom this thesis would have most definitely not been possible. I had the great opportunity to draw upon his vast knowledge and learn from his ability to rapidly understand even the most complex interactions and to always find any weaknesses or strength in a line of argument. He has also always been very supportive and open. One could always get advice and feedback whenever one wanted. I can only admire his efficiency and his rapidness of response to any questions or need for advise I ever had. For all of this and much more, I am immensely grateful to him.

I would also like to thank all the other faculty and postdocs at Gatsby, for making it such a great place to work and help broaden my perspective onto many different areas. Through all the various journal clubs, talks and particularly also tea, there was always ample possibility to discuss and learn a broad variety and get valuable feedback on one's own thoughts. They have all helped create an environment where it was always possible to find someone who was more than willing to take the time to explain and discuss any questions or topics and provide new insights.

At least as important, though, were my fellow peers and grad students in shaping my views and thoughts, as well as providing a great social environment. Out of the many friends I had at Gatsby, I would like to thank some particularly, although each and every one I have known has added in their own special way. Thus I apologise to those I forgot to mention. Foremost I would like to thank Reza Moazzezi, with whom I started neuroscience. I have enjoyed many interesting discussions to dissect and understand new papers and theories with him, as I did with Amit Miller. I would also like to thank Demis Hassabis and Adam Sanborn for numerous discussions on human experimental design. Further, I would like to thank Richard Turner and Misha Ahrens for their fantastic job and passion as TA. Finally, I would also like to thank Charles Blundell, Ross Williamson, Iain Murray, Vinayak Rao, Phillipp Hehrmann and Ritwik Niyogi for being great friends and discussion partners and making the time so enjoyable.

As part of the pleasant and supportive environment at Gatsby, I would of cause like to thank the administrative staff, particularly Alex Boss and Rachel Howes, for their great help in any day to day needs and in general in keeping Gatsby working so smoothly.

Last but not least, I would like to immensely thank my parents who have always provided great support, both during my PhD and long before, as well as my sister, without whom I would have possibly not been in neuroscience or even academia at all.

Contents

List of figures

Chapter 1

Introduction

Many animals and particularly humans show a remarkable ability to learn to adapt to behaviours that are appropriate to changing tasks and environments.

Trying to understand the flexibility and efficiency of learning has been a central concern for many different research fields for centuries, such as psychology, educational research and more recently neuroscience. Equally, it has also been a long standing desire to try and create artificial agents that can emulate and reproduce such flexibility. With the advent of computers, this has turned into a strong focus of research with advances in both machine learning and artificial intelligence. Combining the approaches of studying natural learning systems and constructing artificial ones, cognitive modelling seeks to create such executable algorithms in order to better understand natural ones. By demonstrating that algorithms can reproduce behavioural observations, one can support suggestions of the underlying mechanisms found in natural organisms like humans. This is the approach taken in this thesis, and it aims to make a small but important step towards the understanding of the mechanisms underlying flexibility and rapidness of learning.

Formal empirical research into learning dates back at least to 1885, when Ebbinghaus (1913) first studied the memorisation of lists of separate syllables. It remained a strong emphasis in psychology right through into the 1960s, though thereafter the interest in learning diminished drastically (Shuell, 1986), due to a shift in interest from behaviourist to a cognitive orientation. It never subsided completely though (e.g. Bruner, 1957, 1961; Wittrock, 1974, 1979). Since then, focus in learning in psychology has increasingly been towards modelling, drawing both on empirical research and parallel developments in artificial intelligence. Broadly speaking, these often separated into two classes of models. On the one hand, there are symbolic and logic driven models of production systems, capturing many of the facts known from behavioural human experiments, of which perhaps ACT-R (Anderson, 1996; Anderson et al., 2004) and its variants are the most successful example. On the other hand, there are the more mechanistic, neurally inspired connectionist models, aiming to explain how learning

may be implemented in the brain (e.g. Hinton et al., 1986; McLeod et al., 1998; Elman, 2005). The latter are the ones considered in this thesis, aiming to advance the mechanistic understanding of rapid learning. In addition there have been various attempts to combine the two approaches (e.g. Lebiere and Anderson, 1993; Touretzky and Hinton, 1988; Touretzky, 1990; Rougier et al., 2005).

One key aspect in understanding the flexibility of learning is likely to be the ability to draw upon previous knowledge. I.e. rather than relearning each entire task at the point of change, substantial parts previously learnt may be shared and transferred to the new task. Thus, learning potentially becomes much more rapid, as only an incremental amount of information needs to be acquired. However, one issue that often makes this process of fast sequential learning challenging, particularly in more cognitive style tasks, is the discontinuity that frequently exists. For example, an agent may be faced with sequentially learning 3 different tasks, A, B and C. Perhaps A and C share a lot of common structure and C should benefit from already having learnt A. On the other hand, B may not have commonalities with either A or C, or even have opposing properties. The agent must not forget the at that point irrelevant A while learning B, and retrieve the relevant aspects when learning C. Solving such issues is likely to play an important part in understanding the benefits of sequential learning.

There are, however, at least three different facets to such non-stationary sequential learning, separated both in time scale over which they occur, as well as in the type of information transferred (described in more detail in 2.4). The first facet is the progressive tightening of requirements of the same task. For example, Eckerman et al. (1980) trained pigeons to peck within a small target area by progressively reducing the size of the reinforced area, a technique akin to the common machine learning method of simulated annealing (Duda et al., 2000). A second aspect is the transfer of specific skills or competencies across different tasks to hierarchically build up overall more complex tasks. For example, a rat might learn to lever-press in one stage that can later be used in new and separate tasks. Both these forms of non-stationary learning are often seen over short immediate time scales although they could be used over longer terms, too. The third aspect, on the other hand, typically involves much longer time scales and is commonly referred to as “life long learning” or “learning to learn” (Thrun and Pratt, 1998). In contrast to the previous, there is no specific transfer between tasks. Instead, general biases of the environment, or in statistical terms a prior, is learnt (Baxter, 1997), facilitating later learning. In real world applications, however, these facets seldomly occur in isolation. Particularly the first two methods are often explicitly combined in a further hierarchical way, using an annealing strategy to form individual skills that are then embedded into more complex tasks.

This thesis predominantly focuses on the facet of short term transfer and build-up of competencies to subsequently rapidly adapt to more complex tasks. In order to study these effects systematically, a technique typically called shaping is drawn upon.

The term shaping, commonly attributed to Skinner, refers to a method of "successive approximations" and is now frequently employed when training laboratory animals in complex behavioural tasks. The method suggests breaking down the complex task that the animal should learn into a sequence of sub tasks, each of which on its own is rather simpler, until the animal is eventually able to perform the full complex task.

Despite this wealth of experimental evidence to the benefits of shaping and sequential learning in natural organisms (e.g. Skinner, 1938, 1953; Staddon, 1983; Mazur, 2005), as well as the strong interest of theorists in learning in general, the concept of shaping has so far not found much resonance amongst the various models of human cognitive behaviour, particularly in neurally inspired models of learning. Here, network models are typically initialised in a "naive" state of random connection weights and then trained on a single task in one fell swoop, or perhaps at most on multiple tasks simultaneous. The conditions, implications and advantages of sequential training of partially unconnected tasks on a given network architecture remain mostly unexplored or unclear. One reason for this apparent lack of interest, at least within the class of connectionist models, may be the difficulty in achieving any benefits from sequential learning. It is well known that connectionist models perform poorly when they are training sequentially and thus are learning in a non-stationary environment. When switching from one training regime to the next, connectionist models typically rapidly overwrite and forget the previously learnt information if it is not constantly refreshed. As a consequence, this information is not available to draw upon later, a phenomenon commonly described as *catastrophic interference* (French, 1999; McCloskey and Cohen, 1989). Worse even, learning in a non-naive network can *harm* performance under certain conditions, as network weights no longer have the ideal initial weight distribution, perhaps akin to being "occupied". Among the different types of sequential learning, task sequential learning is the one that triggers these effects most strongly. This is due to the fact that the different tasks on one level of the hierarchy are often fairly independent, such that there is no repeated training of one set of skills for extended periods of time while the network is in the context of another task. As a result, anecdotal observations suggest that a number of researchers have toyed with using shaping to help them train their connectionist models for cognitive tasks (mostly without success), but so far none have made it part of their formal research. In robotics and machine learning, however, a number of attempts have been made to analyse enhanced learning through a sequence of training environments (e.g. Singh, 1992; Saksida et al., 1997; Dorigo and Colombetti, 1998; Abbeel and Ng, 2004; Bengio et al., 2009; Taylor and Stone, 2009).

Since breaking apart a complex task into a multitude of smaller ones automatically establishes a form of hierarchical structure, the work on shaping naturally also has some links to other hierarchical learning approaches increasingly being studied. Of particular interest are the ideas of options (Sutton et al., 1999), described in hierarchical reinforcement learning, and their links to neuroscience (Botvinick et al., 2009). Here, a set of base competencies (policies in the terms of RL) is provided upon which the

more complex learning is based, partially resulting in similar issues as faced in the later stages of shaping, i.e. the selection of appropriate available structure. Indeed, it has previously been suggested that options may arise out of a process of sequential learning such as shaping, although so far they have predominantly been provided either by hand or identified through their statistical properties of the environment, such as temporal or spatial bottleneck states (e.g. Kretchmar et al., 2003; McGovern and Barto, 2001).

As with any aspect, sequential learning and cognitive flexibility can be studied on many different levels depending on the questions asked, each requiring a variety of techniques and methods of analysis but each providing their own insights. Even within the limited realm of computational modelling there is still a broad spectrum of possible questions. For example, Marr (1982) suggested three possible levels of analysis. 1) *Computational theory*, 2) *representations and algorithms* and 3) *hardware implementation*. The most abstract level, the computational level, concerns itself with the question of what the actual goal of an intelligent agent is and the strategy with which it can be achieved. This is often described in purely mathematical notation, with no direct regard of potential implementations. The second level, the representation and algorithmic level, concerns itself with step by step (algorithmic) transformations of input to output representations. As such it must take the general principles and capabilities of the underlying hardware into account, without however looking at the details. In the case of neuroscience, these may include principles such as the fact that computation occurs in large networks of distributed simple computational elements, which is for example captured in the framework of connectionist models. The third, and lowest level, the hardware implementation, has so far been mostly ignored when it comes to modelling of cognitive tasks. Even the most ambitious projects such as the Blue Brain Project have so far only succeeded in scaling "fully biologically plausible" models to the level of a single cortical column, which is by far not enough to study cognitive phenomena. However, hybrid models that combine abstract neural network models with specific aspects of the biological knowledge, such as the dopamine system or genetic variances (e.g. Frank and Hutchison, 2009; Doll et al., 2011), are strarting to emerge.

While all levels are equally important in the overall understanding of the effect of sequential learning on cognitive flexibility, the questions raised in this thesis mostly focus on an intermediary level that bridges the computational theory and the algorithmic level. They aim to elucidate the computational mechanisms required in an algorithmic implementation to support cognitive flexibility, both through the use of sequential learning and in general. As such, a core emphasis is on the computational properties of extensions to standard connectionist models that enable them to take advantage of available structure and hints in the environment. One such architectural support, not unique to sequential learning, is a provision for working memory. It is known that for example standard recurrent neural networks are Turing-complete in their computational power (Siegelmann and Sontag, 1995) and thus would be perfectly capable of dealing with any task, including those requiring memory. However, the inefficiency

and thus the number of training examples required to learn a working memory system would be prohibitively large. The addition of architectural support for a generic short term memory system, however, allows an increase in speed of learning of several orders of magnitude (O'Reilly and Frank, 2005). For this reason, the central aim of this thesis is to identify similar architectural additions that allow connectionist models a further drastic decrease in learning times through the use of additional environmental structure. Instead of extending the architectural support along the lines of known anatomical facts, however, the approach here is to first identify basic computational needs, which in future can hopefully be mapped onto neural systems.

This emphasis on an intermediary level between computational theory and algorithmic implementation is naturally reflected in the choice of modelling frameworks used to address the key questions raised. With the focus on the computational properties of the architectural support necessary to benefit from sequential learning, models clearly had to be based on the general architecture, i.e. on connectionist models under investigation, or at least provide a trivial transform to the latter. However, given the emphasis on the computational principles of these architectural additions, it was important to choose the most abstract models that still provided the necessary link to the mechanistic implementation in order to give the purest and simplest model possible. This reductionist simplicity naturally came at the expense of biological detail and plausibility, although all results should transfer relatively easily to the more biologically inspired models. Nevertheless, although the emphasis was towards the mechanistic level, key computational inspirations originated from both abstract Bayesian models as well as reinforcement learning algorithms, providing a link between both forms of modelling in the context of task sequential learning.

As a result, this thesis provides one of the first systematic approaches to modelling short term task sequential learning in a neurally inspired cognitive setting.

The problem of modelling sequential learning is approached in the following way: To begin with chapter 2 reviews four lines of research that are relevant to the models presented in later chapters. First, in highlighting the most important components in studying models of shaping in cognitive tasks, the general concept of shaping and some of the main properties and findings as known from psychology and behavioural testing are presented. In this context, some of the prominent examples and algorithms of using shaping to improve learning in robotics and machine learning are reviewed. Secondly, a more in-depth review of catastrophic forgetting is provided to discuss the core limitations and difficulties of connectionist models in sequential learning. Thirdly, some of the relevant properties for modelling of two of the brain areas most closely linked with flexible behaviour and learning, the prefrontal cortex and the basal ganglia are presented. These areas have also been suggested as the biological substrate for the models considered, the class of "gated working memory models", the fourth line of research reviewed. They comprise a class of emerging, neurally inspired models that

capture, in a simple and abstract way, many of the important properties one is likely to encounter when attempting to explain high-level cognitive behaviour and learning.

Given the broad overview of the concepts and components necessary in modelling shaping, chapter 3 focuses in greater detail on the specific models used in this thesis, the Prefrontal cortex, Basal ganglia Working Memory model, the Long Short-Term Memory model (used in chapter 4) and the gated bi-linear model (used in chapter 6). Additional aspects of the models, specific to the modelling presented, are also described in later chapters.

Chapter 4 presents a first connectionist style model of shaping, based upon the Long Short-Term Memory model. In this framework the thesis explores the benefits and downsides of shaping, and compares learning in the typical complete task-at-once, or concurrent, training to the more staged and sequential training of shaping. It analyses the speed of learning, as well as forms of abstraction and generalisation as a measure of cognitive flexibility. Importantly, the need for specialised mechanisms or architectural support in neural networks is discussed and an automated algorithm to overcome many of the issues is presented. This work has mostly been published in Krueger and Dayan (2009).

This chapter illustrates the issue using the 12-AX task, which was specifically designed to highlight the features of the class of gated working memory models. However, it was originally invented purely for the purpose of modelling and so far there are no human or animal data as to how it is learnt. Thus, in a switch of gears, chapter 5 presents human behavioural data on learning in a setting as close as possible to that encountered by the models, showing that the task is indeed suitable for human experimentation. This provides hints as to how well humans can benefit from the shaping protocols explored in chapter 4.

Finally, chapter 6, returns to modelling. It looks at cognitive flexibility from a slightly different angle. Instead of focusing on shaping, it considers the question of the separation of rules from symbols, i.e. the ability to apply the same set of abstract rules on various different set of stimuli or symbols, without needing to relearn the abstract rules from scratch. This is another important aspect of the ability of humans to adapt to new situations rapidly. Key to this form of flexibility is the model's ability for indirection and the use of pointers, a concept that is not commonly associated with connectionist style models. However, this chapter shows, that with the architectural mechanisms provided by gated working memory, the necessary support to allow this form of flexibility is readily available and exploitable.

Overall, this thesis presents models of how shaping can be made to work in the class of gated working memory models, a powerful and popular class for modelling flexible cognitive behaviour. It highlights some of the key benefits of shaping, such as the ability to learn more rapidly, to generalise better and to form abstraction, that may result

from differences in representations. These are all properties commonly found in human performance and they are important to understand in the endeavour of deciphering this remarkable performance. Furthermore, two main aspects emerge from the models, suggesting that agents which show cognitive flexibility and are successful in exploiting sequential learning, are likely to require mechanisms of indirection and rapid targeted structural segregation. No direct and immediately measurable experimental predictions emerge at the abstract level of modelling presented here, but the models do lead on to suggesting specific computational properties that have not yet been strongly explored in neuroscience. The models thus provide a strong conceptual theory which can potentially guide future experiments in understanding the neurobiological mechanisms underlying flexible cognition. In addition to the experimental implications, this thesis reaffirms that task sequential learning has negative effects without appropriate architectural support from the models, explaining a) why shaping has commonly been neglected in connectionist models of cognition, as well as b) why taking task sequential learning into account is eventually essential in understanding the neural architecture of flexible cognition. Nevertheless, the human behavioural experiments also show that even with the help of shaping, current neurologically inspired models could still not match human performance on rule based tasks such as the 12-AX, suggesting that substantial further research will be necessary in this field.

Chapter 2

Sequential learning and flexibility

Cognitive flexibility and the ability to learn progressively more complex tasks with increasing ease over time have fascinated scientists for centuries and are studied across psychology, neuroscience, machine learning and educational research. As such there is a wide range of literature touching on many important aspects of what is covered in this thesis, either directly or indirectly. This chapter thus aims to give a brief overview of some of the most relevant topics.

It starts by covering some of the general properties of the technique of shaping and its applications in psychology and in behavioural experimentation. Next, it presents a number of related models and algorithms that have been proposed to demonstrate the advantage of learning in steps. Although this thesis predominantly focuses on task sequential learning, for a broad overview it also briefly includes models covering the two other facets, learning in progressive steps and learning to learn. Unlike the present models, the majority of these models originated as pure machine learning or robotics applications, especially since the new emphasis on transfer learning in recent years in reinforcement learning (Taylor and Stone, 2009). In contrast the models presented here aim towards explaining natural cognitive behaviour, although they aren't biologically detailed. For this they use the group of gated working memory models, a class of models designed to capture a number of core computational principles believed to sub-serve flexible natural behaviour. As such it finally presents a review of the class of gated working memory models together with the neuroscientific background they are based upon.

2.1 Terminology and definitions

Throughout this thesis, specific terminology is used to explain concepts, methods and results. As such, it is important to provide a clear description of this terminology, which partly has a rich and complex history of definitions and applications. While

the terminology used in this thesis is closely related to existing definitions, it also has its own specific meaning described, in more detail here. Where appropriate, further explanation and discussion of concepts are also found in later chapters where they are used.

Cognitive flexibility. Cognitive flexibility refers to the ability of a system to adapt to a variety of tasks and environments. One aspect of cognitive flexibility is the ability to respond differentially to stimuli in different contexts, i.e to switch response mappings depending on the current context. A second aspect to flexibility is the ability to adapt to entirely new tasks and environments that have never been encountered before in this form. Part of the definition of flexibility is the rapidness with which these adaptations occur. If an agent requires for adapting to a new task a large number of training examples relative to the underlying amount of information necessary for the change, then it is less flexible. Thus the efficiency with which a learning agent uses the information provided by the training examples to adapt is a key measure of flexibility.

Working memory. The terminology of working memory has a long history going back at least to Hebb (1949). In subsection 2.5.1 a detailed description of the general concept is given as is used in the wider literature of psychology and neuroscience. In the context of this thesis, however, working memory has a more abstract and less specific definition than is commonly used. Here, working memory refers to a general form of short term memory, where the information is encoded in the activity of neurons rather than in the synaptic weights. It provides an internal context to the system that allows biasing of responses to current stimuli based on events in the past, transforming it from an agent capable of only simple stimulus response behaviour to one capable of complex cognitive and flexible behaviours. In the computational sense, it provides an augmentation of the state vector beyond the immediate sensory input and thus allows an agent to act conditionally on the broader temporally extended context. Due to this abstract and general definition, the concept of working memory as referred to in this thesis does not directly inherit many of the well known properties of working memory, such as for example its strict capacity constraint or attentional limitations. It is more akin to the idea of sustained firing of neurons in animals during delay periods to signal past stimuli.

Sequential learning. In general, sequential learning refers to a paradigm of learning in which the ordering of stimulus presentations is an important aspect of learning.

There are at least two broad variants of this type of paradigm. First, the task itself can be dependent on sequential information, i.e. an agent cannot perform the task correctly without considering the ordering of the stimuli, since the stimulus presentations are not independent. Second, learning of the task can take place in a non-stationary environment, in which a sequence of changes to the environment occurs during the learning period, even though the final task may not require sequential information. Within the latter category (assuming the sequence of changes is beneficial to the overall learning), there are further distinct types of sequential learning, including *task sequential learning*, *learning in progressive steps* and *learning to learn*, which are each described in more detail in the context of computational models in section 2.4. While all of these fall under the definition of sequential learning, in this thesis sequential learning generally refers to task sequential learning (unless stated otherwise), as this aspect is the main focus of the thesis. Specifically, it refers to learning a sequence of distinct tasks one after the other, some of which may depend on others, while others may be entirely independent and are only linked in later tasks that depend on both. There is no overlapping training of tasks.

Shaping. The terminology of shaping is generally attributed first to Skinner and is defined as "teaching organisms to perform complex behaviours through successive approximations". Since shaping is most commonly referred to in animal training, its meaning has generally been closely linked with the specific type of sequential learning found there. In this thesis, however, that form of sequential learning is referred to as "learning in progressive steps" and is not entirely equivalent to the use of shaping here. Instead, this thesis uses the term shaping more in line with the definition of Erez and Smart (2008), where it is defined in very general terms of "as any trajectory in task space, leading from simple tasks to harder ones with the objective to facilitate learning". In this abstract context, the definition again does not invoke many of the specific psychological phenomena often attributed to shaping, such as fading or stimulus control.

Connectionist models. The connectionist model is a neurobiological model of information processing. It draws an analogy to the central nervous system, in which neurons activate and inhibit each other in complex networks. In this model information processing involves large numbers of units stimulating or inhibiting each other through networks of connections.

Architectural support. The outcome of training a (connectionist) model depends on a combination of the training examples provided by the environment, the learning rule and the underlying model. In principle, a sufficiently general model and learning algorithm can be capable of learning any given task, but the number of training samples required for learning may be unacceptably large. To overcome this problem, architectural support provides additions to the model or learning algorithm to facilitate efficient learning from specific aspects of the environment. It provides general built-in mechanisms to enable it to take advantage of specific structures in the environment, such that the learning algorithm does not need to exploit it by itself. As such, architectural support refers to prior information built in as components of the model to restrict or bias the learning to more efficiently learn a task.

Biological plausibility. In general, the concept of biological plausibility refers to models that are consistent with existing biological knowledge. It is commonly used evaluate models and guide choices in the modelling framework based on what is thought to be most biologically plausible. However, the term biological plausibility is nevertheless often ill defined and controversial. Since it is extremely difficult to generate a model that is consistent with all existing biological knowledge, the term biological plausibility must therefore be seen in context of a specific subset of biological knowledge. The models considered in this thesis are only biologically plausible in the most abstract sense, although the biological plausibility that is invoked does provide considerable constraints on the choice of models, for example ruling out techniques such as purely Bayesian models. The biological plausibility considered here is that of the connectionist framework, which requires that the predominant computation occurs in a distributed network of linear combinations of network activations. Some of the comparison network models presented take the biological plausibility further by aiming to respect anatomical knowledge of connectivity patterns between brain areas (e.g. Hazy et al.), although much of the biological knowledge remains abstracted away even in the "more biological models".

Resource allocation. Resource allocation refers to the use of an external entity to determine whether a specific resource should participate in a task or not. It is a term not commonly used in neuroscience, although it has been used both in the attentional literature, as well as to refer to allocation of neural resources as tasks become more difficult (Mikels and Reuter-Lorenz, 2004; Banich, 1998). In the context of this thesis, it refers to a form of structural modularisation of a connectionist network. A resource is one or more distinct parts of the network (working memory modules) that are allocated to a task for learning. Learning is thus not only a function of input and internal

representations, but also a function of whether their neurons have been allocated to participate in learning a task. Specifically in the context of the model of Long-Short Term Memory, it refers to the whole scale enablement and disablement of the learning rate of individual working memory modules to allocate a subset of modules in which learning occurs. Thus modules that are not allocated for learning are safe from being overwritten by new knowledge. The decision of which modules are allocated for learning occurs over the temporal extension of a task. In more biologically plausible networks, this would refer to differential metaplasticity (i.e. the change of the susceptibility to plasticity independently per neuron, or computationally an updatable per neuron learning rate) of groups of neurons to limit learning to a varying subset of the network at any given time. This is in contrast to metaplasticity for homoeostatic purposes.

Associative learning. Associative learning is a form of statistical learning, i.e. the learning of associations or contingencies that have a higher probability of occurring in the environment. Due to its statistical nature, this form of learning is usually gradual and continuous. Its terminology is extended into the operant space, in which state actions can get associated with expected future reward through repeated exposure.

Hypothesis testing. Learning through hypothesis testing is a more explicit form of learning than associative learning. At any given time, an agent has a complete hypothesis of the task instructions or world model. The agent then validates the predictions of its hypothesis against the samples it receives from the environment. Whenever the environmental stimuli conflict with the predictions of the hypothesis, the hypothesis is partially updated or replaced in full. Instead of gradually learning, hypothesis learns in steps. If a good probabilistic distribution on likely hypothesis exist a priori, this form of learning can be more efficient than associative learning. Furthermore, in the context of working memory tasks, a hypothesis can include stimuli to be ignored and can thus potentially cope with larger temporal spans of distractors, without rapidly escalating training times.

2.2 A brief overview to the background of shaping

Shaping is a technique for teaching organisms to perform complex behaviours through "successive approximations". Although similar ideas have long been used for teaching and training, their formal study is typically attributed to Skinner. One of the first classical accounts was a description of how he taught his lab rat Pliny the behaviour of *"pulling a string to obtain a marble from a rack, picking the marble up with the*

fore- paws, lifting it to the top of the tube, and dropping it inside. Every step in the process had to be worked-out through a series of approximations, since the component responses were not in the original repertoire of the rat.“ (Skinner, 1938). Shaping has been referred to both in terms of learning in a continuous progressive steps (“successive approximations”) as well as for the hierarchical skill based learning (“steps” and “components”). More recently, Erez and Smart (2008) generalised the definition of shaping further, declaring it in computational terms as any trajectory in task space, leading from simple tasks to harder ones with the objective to facilitate learning.

2.2.1 Theories of shaping

The basis of shaping. Crucial to shaping is the decomposition of tasks into appropriate, sufficiently small steps to ensure that at each stage the reinforced action is emitted spontaneously (with respect to the agent's current training status) with an adequately high probability. Shaping originally builds upon the ideas of operant conditioning, in which one behaviour or response is selectively reinforced, drastically increasing the probability or frequency of the animal emitting this behaviour (Thorndike and Woodworth, 1901). As such, it relies on the desired behaviour to be spontaneously emitted, in order to be able to reward it. However, the probability of naturally performing complex behaviours (such as the one shown by Pliny the rat) is often too low to successfully rely on trial-and-error of pure operant conditioning, even if the animals are ultimately able to execute it without problems. For this reason, intermediate reinforcement is required to facilitate this process.

In the traditional view, shaping is split into two important contingencies: Selective reinforcement and extinction (Martin et al., 1999). The selective reinforcer progresses over time, implementing the “successive approximations”, guiding action selection. The two main effects of extinction are to eliminate earlier parts of the shaping procedure that are now redundant, and to enhance variability to increase the likelihood that the new reinforced response is emitted.

Selectivity and contact. There are two important properties of a shaping protocol, “selectivity” and “contact”. Both refer to how a subject is rewarded in the course of shaping (Platt, 1973). In a shaping protocol based on selectivity, the subject is required to achieve a specified absolute performance level in order to receive a reward in a given sub-task. In other words, selectivity determines how close to the target behaviour the subject has to perform to be rewarded. For example, if the final aim is to achieve a behaviour requiring the subject to wait X seconds, selectivity determines the minimum wait time before a reward is given in the sub-tasks. A high selectivity forces the subject to show a good approximation. However, the reward rate might be very low. On the other hand a contact based shaping protocol uses a fixed frequency with which an agent receives a reward. For example, the best 10% of approximations

may be rewarded, irrespective of how good they are in absolute terms, guaranteeing a specific reward rate. Hybrid approaches have also been used, in which for example the selectivity criterion is made less stringent if no sufficient contact is achieved (Pear and Legris, 1987). The property of selectivity has traditionally been more common (Midgley et al., 1989) and is inherent in the hierarchical modular shaping.

Behavioural units. One important outcome of shaping is to create new "behavioural units" like lever-pressing (Schwartz and Robbins, 1995). A behavioural unit is usually a unique stereotypical behaviour. While the behaviour might display a considerable initial variability, it should be optimised during training to the minimal possible response that still results in the reward. As superfluous elements of the behaviour are extinguished, the overall variability is reduced. One important consequence of forming a behavioural unit is that it extinguishes and, more importantly, can be re-instantiated again as a whole. Prior to forming a unit, each behaviour is extinguished independently. Thus, a form of modularisation occurs.

Pavlovian influences on operant shaping. In the operant theory of shaping, reinforcement simply acts to strengthen the stimulus - response mapping, increasing the likelihood of eliciting the response independent of the reinforcer. However, there are also accounts of a Pavlovian influence (Bindra, 1972, 1974), according to which the association occurs between the stimulus and the reinforcer directly. The animal's response is then determined by the current motivational state and the conditioned incentive properties of the stimulus. For example, if a pigeon is rewarded with food, it will associate the stimulus with food, which results in an approach action together with pecking. Due to the direct link between reward and action, the type of reinforcer plays an important role in determining which responses are easy or even possible to shape, unlike in operant conditioning, where only the current value of the reward to the subject is important.

The Pavlovian effects of shaping appear to introduce some clear limitations to animal shaping. Breland and Breland (1961) provided some striking examples of failure cases. In one of these, they tried to teach a raccoon to deposit a coin in a piggy bank. Over time, the raccoon actually started to perform worse with progressive stages. Instead of depositing the coin in the box, for which it was rewarded, the raccoon could not let go of the coin and started playing with the coin instead. In an explanation similar to the model of Bindra (1972), they suggested that the raccoon seemed to associate the coin directly with the food reinforcer, letting the innate reaction to food overwrite the operantly expected behaviour.

Shaping-history dependent behaviour. One interesting question arising, particularly with respect to the difference between operant and Pavlovian influences, is whether

the exact nature of the shaping protocol influences the final behaviour itself or just the speed of learning. If the properties of the final behaviour differ depending on the shaping protocol chosen, manipulating the shaping protocol might help reveal mechanisms and representations involved in transferal during shaping. Stokes and Balsam (1991) conducted a set of experiments, testing the effect of different shaping approximations. Two groups of rats were trained to lever-press, with different reinforced actions used to guide the rats to the lever in the two groups. For one group, the approximations were based on rearing, while in the other group progressive nosing responses were reinforced. In the resulting lever-pressing, the two groups differed. The rearing group continued to show rearing behaviour during lever pressing, while the nosing group did not. Furthermore, in a separate experiment, the behaviour did not fall back to a single stereotypical behaviour associated with the reinforcer with extensive training. For changes in both early and late stages of shaping, the resulting behaviour differed even after substantial periods of exercising the final task. This result would not have been predicted by the Pavlovian parts of shaping, as the innate behaviours remained unchanged. These results are particularly interesting with respect to the experiments presented here, as they do not model the Pavlovian aspects and predict a differential behaviour based on shaping.

Shaping applied to cognitive tasks. The pure behaviourist view of shaping often quickly reaches its limits when applied to complex cognitive, non-physical behaviour, as is typical in human educational teaching (Gagné, 1962). One reason is that stimulus-response contingencies are less important in the context of manipulating memories and other hidden internal representations required for learning concepts, rules, principles, intellectual skills, and other cognitive strategies. Moreover, the steps of learning are often not representable in a simple linear sequence. They rather grow out of the idea that any particular, higher-order intellectual skill is based on several lower-order intellectual skills, and that each of these in turn is based on even lower-order skills. Gagné (1970) took the core idea of behaviourist shaping, that of "successive approximations", and extended it into the cognitive domain. This came to be known as cognitive behaviourism.

For example, in their method of "hierarchical task analysis", Gagné (1970) proposed that in order to efficiently teach an overall complex task, such as solving algebraic problems, the task has to be broken down into a hierarchy of sub-skills. Each sub-skill is a skill in its own right for which the method can again be used to learn, resulting in a hierarchical recursive procedure. While prior to teaching a higher level skill all of its sub-skills need to be learnt, teaching does not need to be sequential with in one level. This method thus provided an systematic way of constructing a teaching schedule.

2.2.2 Applications and considerations of shaping

The art of shaping protocols. Despite the intense study of shaping, it often remains unclear what properties make a shaping protocol particularly efficient, and under which conditions. The focus of behaviourists in teaching had mostly been on "how", rather than on "what" to learn, leaving it open how to define "successive" or "approximation" (Case and Bereiter, 1984). While in simple motor behaviour, "successive" can sometimes be translated into its physical distance to the target, the question is less clear in cognitive tasks. This left e.g. Midgley et al. (1989) to declare that *"its practise could fairly be said to be more an art than a science"*. Trying to rectify the lack of rigour, several studies aimed to define shaping in algorithmic terms (Midgley et al., 1989; Platt, 1973; Pear and Legris, 1987). Specifying the shaping schedule in terms of a few rules (i.e. simple algorithms) made it possible to verify that these captured the essence of shaping and to compare different strategies in their effectiveness. One such simple rule was "Always reinforce the best 20% of ongoing responding", which Platt (1973) termed "percentile reinforcement" schedule. Another was "to reinforce all responses achieving a minimum property, progressively raising the minimum", which each pick up the properties of contact and selectivity.

Practical considerations of shaping. Neither the limitations of operant shaping, nor the lack of clear theories to create shaping protocols, have hindered its adoption for practical animal training. Shaping continues to be widely employed, as a key way of teaching animals to perform the complex tasks researchers are interested in. The vast majority of behavioural studies involve some form of shaping, even though it is rarely the focus of the study itself.

The speed with which shaping allows animals to learn varies greatly, depending both on the species trained and the task involved. While in many rat and mouse experiments training times lie in the range of several weeks and tens of sessions, monkey experiments often require much longer, ranging into the time frame of a year (e.g. Shima et al., 2007; Averbeck et al., 2006), although it heavily depends on the complexity of the task. Under some conditions though, shaping in animals can be much more rapid, too.

It was originally believed that hand shaping may be more efficient than algorithmic shaping, based on observations by Skinner in 1943 (Peterson, 2004). In these experiments, Skinner was able to teach a wild pigeon to push a ball into a hole within only a few minutes by guiding it with food, much faster than in previous experiments in which he had shaped the animals through manipulations in the animals environment. Later however Midgley et al. (1989) showed that this depended on the nature of the algorithmic shaping protocol used. In this study, they were able to develop a shaping protocol that was just as efficient as hand shaping in training a rat to push a ball into a hole. The shaping protocol they used closely resembled what they thought was the

essence of hand shaping, resulting in an algorithm controlling both for selectivity and contact. Although the base shaping protocol was selective, the algorithm automatically moved back to simpler approximations if the reward rate dropped. It also allowed the subjects to skip stages of shaping if they emitted the correct behaviour of later stages early.

Metrics for successful shaping. Shaping and in general learning in stages can provide a variety of different ways in which an agent's learning can benefit. As such a variety of different metrics exist to evaluate how successful either a shaping protocol or a learning architecture have been in improving performance. Taylor and Stone (2009) listed 5 different useful metrics for transfer learning in reinforcement learning, which equally apply to shaping approaches: Jump start, which is the performance at the beginning of the target task or instantaneous transfer, asymptotic performance, total reward earned during training, transfer ratio and the training time to reach a specific performance threshold. In addition, they distinguish between a total time scenario and a target task scenario. In the former, performance is evaluated including the full shaping sequence, which is particularly relevant when shaping occurs specifically to train a single task, while the later only includes the final task when calculating the metric, ignoring time spent during the sequence of shaping tasks.

2.3 Catastrophic interference in sequential learning

It has long been known that connectionist models have shown disruption of older knowledge when trained on new elements (e.g. Carpenter and Grossberg, 1986; Hinton et al., 1986). It is part of the general plasticity-stability dilemma (Grossberg, 1980), which highlights the problem of a learning system to remain plastic in response to significant new events, but also remain stable in response to irrelevant events. In a detailed analysis of the issue, McCloskey and Cohen (1989) coined the term "catastrophic interference" to describe this problem. They suggested the issue for connectionist models might even be *fundamental*, i.e. without a solution, disputing earlier accounts (e.g. Hinton et al., 1986), that the problem may be "mild" and thus neglected. The term refers to that in connectionist networks new learning completely disrupts or erases previously learnt information, rather than a gradual degradation as new tasks get learnt, as is more common in humans (Braun et al., 2001). McCloskey and Cohen (1989) used the example of learning arithmetic tables, sequentially training the one-times table, and then the two-times table, to show this rather abrupt and complete forgetting.

Since then, a significant amount of research has gone into attempting to overcome the catastrophic interference, alleviate it, or indeed prove its fundamental status. Early work by Kortge (1990) aimed to solve the issue by reducing the overlap of inputs and limit the number of weights updated during learning ("we would like to blame

just those active units which were responsible for the error"). These ideas resulted in "semi-distributed" representations, somewhere between localist and fully distributed representations. The methods of how to achieve such reduced overlap, however, varied from e.g. input encoding (Murre, 1992; French, 1994) to training method (McRae and Hetherington, 1993).

Separately, the idea of a dual-net or bi-modal structure had emerged (French, 1992) and has become more popular since. Here, the network was separated into two parts, each of which had a different learning rate and thus lay on a different part of the stability - plasticity spectrum. These ideas were partially picked up by McClelland et al. (1995) and extended with many insights from systems neuroscience to create the Complementary Learning System (CLS) (O'Reilly and Norman, 2002; O'Reilly and Rudy, 2001; Norman and O'Reilly, 2003). In this model, the network is split into two parts, one of which is responsible for rapid and automatic memorisation of patterns, while the other is a longer term, more integrative system. Based on what is known about human learning and memory, it was proposed that these two parts of the network may correspond to different brain regions that are involved in memory formation, with the former representing the hippocampus and the latter the neocortex. The hippocampus played the role of the rapid and automatic memorisation of patterns and the neocortex the role of the longer term more, integrative system. However, in its original formulation, the CLS showed unacceptably high rates of forgetting in non-stationary environments, when training examples were no longer shown (Norman et al., 2005) and it required a secondary off-line learning mechanism to overcome.

Hence, while catastrophic interference is inherent in connectionist networks, a variety of methods provide at least partial solutions to the issue by introducing different architectural sophistications.

2.4 Enhanced learning in artificial systems drawing upon prior knowledge

Despite the great advances in robotics and machine learning, there is still a large gap in performance relative to natural learning systems in moderately to complex tasks. It is recognised that one of the key shortcomings of artificial learners is that they frequently initiate learning from a tabula rasa (Bernstein, 1999). As such, a number of approaches to improve learning using incorporation of prior knowledge through sequential learning have been studied and developed. Especially in recent years, a renewed interest in transfer learning in reinforcement learning has emerged. Nevertheless, many of the approaches so far have been mostly independent proofs of principles, demonstrating the feasibility of improved learning under various circumstances. Furthermore, approaches have differed in many respects, covering various different learning architectures, different levels of human intervention as well as different types of utilising prior knowledge.

The vast majority have purely focused on the engineering and computational level, although in a few notable examples, shaping has also been used in modelling and understanding human or natural behaviour.

This overview of the different previous approaches groups them according to their use of prior knowledge and the resulting properties of the sequence of tasks.

2.4.1 Task sequential learning

In the task sequential facet of shaping, each stage of shaping constitutes a separate task and provides a simple way of learning one specific aspect of the environment or final task. Latter tasks can then depend on the skills acquired in earlier shaping tasks, creating a hierarchical construction. Key to this method of shaping is for the teacher to analyse the final task and decompose it into its components, ensuring the ordering of shaping satisfies all dependencies (Gagné, 1962; Dorigo and Colombetti, 1998). Thus, shaping provides a way to learn a hierarchical task in a bottom-up way. As the transfer between tasks occurs mostly in form of distinct competencies, nearly all approaches allow for some form of modularity in the architecture to prevent interference. Given that modularity is often enforced by the model itself rather than emerging through learning, it has been referred to as "structural modularity" (Gullapalli, 1992).

One such approach has been the model of Asadi and Huber (2007). As an abstract reinforcement learning model, it drew upon the hierarchical framework of options (Sutton et al., 1999) modelling the environment as a semi-Markov Decision Process (SMDP). For each task in the sequence, once a policy had been learnt for it, sub-goals were extracted from the system model, and corresponding sub-goal skills were learnt offline. These were then explicitly added to the repertoire of options for subsequent tasks. To further profit from the transfer between tasks, the model also constructed a hierarchical state space representation to compact the underlying state space into a more abstract representation. They overall showed that their model could learn more rapidly to navigate in a virtual world than an equivalent model without shaping.

In a different approach, Gullapalli (1992) demonstrated another modular approach to shaping. Here, he simulated a counting task, which he broke into a number of hierarchically organised sub-tasks, each of which he trained the network on independently. The architecture represented an explicit two level controller, in which the actions of the higher level were to engage lower level modules. Shaping consisted of training the higher level controller on the complex task only once the lower level was fully trained. Again, he showed that without the help of shaping, the controller did not learn at all, in contrast to the structurally shaped network.

In yet another approach Urzelai et al. (1998) demonstrated improved learning performance in a robotic navigation task through sequential decomposition and training, called "behaviour analysis and training" (BAT) (Dorigo and Colombetti, 1998). This

model again explicitly employed a modular structure, with each module containing a full, feed forward based action controller. These were in turn trained on individual sub-tasks. Although Urzelai et al. (1998) explicitly referred to their model as robot shaping, its use of genetic algorithms for learning prevented a single agent learning from a continuous sequence of tasks.

While the above models drew upon core ideas of task sequential learning, i.e. to decompose a complex task into sub-modules and learn these in a bottom-up fashion, they did not directly address one of its striking difficulties. By manually assigning tasks to modules these models did not look into how a sequence of tasks might be automatically decomposed into an appropriate modular structure.

In a model more explicitly designed to decompose a sequence of tasks, Singh (1992) introduced the algorithm "compositional Q-Learning (CQ-L)". This algorithm heavily drew upon the ideas of a class of models called "Mixture of Experts" (Jacobs et al., 1991b) that formed an important foundation for automatic modularisation.

The Mixture of Experts supervised learning model consisted of a number of independent neural networks, the experts, which were arbitrated through a gating network to form a common output. For learning, the supervised error signal was differentially applied to the expert networks, proportional to the choice probabilities calculated by the gating network. With a selective gating network, adaptation through learning is therefore limited to only a few responsible experts at any given time, inducing a modular structure. Singh extended this idea into a temporally extended (stateful) setting. Here, gating occurred between extended sequential decision policies of their own. Each expert was replaced by a full Q-learning module. The gating network was trained to maximise the likelihood of achieving the desired Q-value. Due to the temporally extended nature of its tasks, and their overlap in instantaneous input space, the gating network was extended to include input from a specific task signal. The key result was that for composite tasks, consisting of concatenations of multiple elementary tasks, C-QL could learn overall faster than a simple Q-learning module alone, due to its gating network and transfer between tasks. While the elementary tasks actually learnt slower than with plane Q-learning, the reuse of the elementary Q-learning modules in the composite task allowed it to learn considerably faster than learning the composite task from scratch. One important problem with respect to employing this model in a shaping setting was, that for the gating network and Q-modules to learn correctly, both elementary and composite tasks had to be randomly mixed during training.

To extend this model to the task sequential setting of shaping, Singh no longer trained on randomly mixed samples of both elementary and composite tasks, but trained the elementary task to completion before exposing the model to the full composite task. The model correctly and efficiently learnt the composite task, separating out the secondary elementary task into the additional Q-learning module, despite never having been trained on that elementary task alone. However, in this setting the gating net-

work was no longer successful at correctly learning responsibilities. An additional intervention was necessary to prevent the network from unlearning the first elementary task during the long runs of identical tasks. This was achieved by disabling learning in the Q-learning module trained on the first task once the switch to the composite task occurred, thus loosing its ability to automatically assign modularisation.

2.4.2 Learning in progressive steps

The second approach to benefiting from learning in steps focuses more upon intra rather than inter-task transfer. Instead of building up distinct units of competencies that are transferable to new tasks, this approach aims to scale the difficulty of a single task along a parametric axis. As each stage of shaping is a variation of the previous task, typically a strict super set, it does not face the same issue of interference as the task sequential approach. These methods, therefore, do not extensively employ forms of modularisation. Nevertheless, it is perhaps the version most closely related to the use of shaping in behavioural and psychological work and can be seen as a direct interpretation of "successive approximations". It is also likely to be the most widely applied and studied approach, although again there is a substantial variation in the properties of approximation and scaling, ranging from altering rewards, over action space to physical properties of the environment.

One classic example of this approach, although probably also the most distant from the behavioural aspects of shaping, is the widely used technique of simulated annealing for optimisation (Duda et al., 2000). By scaling the "temperature" parameter in its optimisation function, it decreases the chance of getting stuck in a local optimum, increasing the chance to find the global one. Unlike methods of shaping, though, the unguided global property scaled in simulated annealing does not explicitly direct learning towards the goal.

Environmental shaping. One possible dimension along which difficulty can be scaled is the environment or task itself, i.e. the underlying model of the environment can be altered in order to reduce its complexity.

In a classic task of motor control learning, the pole balancing task (e.g. Barto et al., 1988), Selfridge et al. (1985) have shown that training on a simpler task first can speed up learning a harder task later. In this task, a robot needed to balance a pole by moving the cart on which the pole was balanced back and forth. Here, the scaling of the environment was to learn to balance a pole with a lighter mass, which is easier to learn due to the available force to mass ratio. Even when combining the training times of both tasks, they were less than learning the harder task in one go.

In a model unrelated to RL, Bengio et al. (2009) have recently studied shaping in training multi-layer neural networks to classify various shape images into rectangles

and ellipses. Here, the successive approximations consisted in initially only training on examples of the simpler sub-class of square and circle images, before showing the full set. Bengio et al. illuminated some of the potential computational principles behind the positive effects of pre-training. They postulated two possible reasons for faster learning: One is that the pre-training example learning efficiency increases if the difficult ones are initially ignored, where they would not yet benefit learning. The second hypothesis is that the intermediate examples allowed the network to find more optimal local minima in the error manifold of parameter space.

In a model designed to study natural language learning in humans, Elman (1993) looked into the effect of "starting small" in a connectionist model of language, realising a concept described by Newport (1988, 1990) in terms of "Less is More". In an artificial grammar, designed to incorporate many of the features believed to make language acquisition difficult, he analysed the effect of training the network with a simpler, restricted set of training examples first, only later exposing the network to training example of the full complexity of the language. In contrast to a common belief that connectionist networks are best trained on the full data set for some problems (Harris, 1992), Elman found that learning was only successful when training in stages.

Reward shaping. In reward shaping the external environment as well as the agent is left unchanged from the original task to be trained. However, a teacher inserts additional reward or punishment as a guidance to the overall task goal. Due to the additional rewards introduced, reward shaping induces a form of temporal scaling in that the expected time between rewards is reduced, alleviating the issue of temporal credit assignment. Reward shaping is one of the most common forms of shaping employed to help artificial agents to learn faster. Not all models of reward shaping, however, provide a dynamic succession of tasks, instead altering the reward function in a static way (Laud and DeJong, 2002).

As an example of reward shaping, Gullapalli (1992, 1997) demonstrated the learning of a robot arm to press buttons on a calculator with a realistically simulated environment. While the feed-forward neural network controller learnt to press the correct button out of three when shaping was applied, the unshaped network did not learn even after very long training periods. Shaping broke the overall task into small consecutive sub-parts of the form "lift hand from keypad" or "Move the fingertip towards target key" which were each trained until the robot achieved acceptable performance on them.

In another example of reward shaping, Randløv and Alstrøm (1998) demonstrated learning of balancing and riding a bike with a Sarsa(λ) model (Rummery and Niranjan, 1994) by introducing additional rewards, accelerating learning significantly.

In a more elaborate model of shaping, Dorigo and Colombetti (1993, 1998) defined it in form of an active teacher agent, which they termed "Reinforcement Program" (RP).

This teaching agent, which provided additional rewards, was in fact embedded in the same environment as the learning agent, having to react and adapt directly to the behaviour of the agent. As such, the task of programming the teacher can be just as complex as the original task. Note however that in low-level robotics, the complexity often comes from control of the actuators, which the teacher would not need. Dorigo and Colombetti also suggested that the teacher may possess more or better sensors to reduce its complexity. They did speculate, however, that the teacher need not necessarily be artificial, but its role could be played by e.g. a human. Nevertheless, they claimed that the temporal precision and speed of a human made it hard to match the behaviour of a robot efficiently.

In contradiction to this claim, Saksida et al. (1997) used a real human teacher to demonstrate the effectiveness of their shaping procedure in some of their experiments in robotic shaping. Like in Dorigo and Colombetti (1993), the teacher in Saksida et al. took a rather active role in explicitly trying to guide the agent to the desired behaviour. Overall, however, Saksida et al. focused on modelling psychological and behavioural properties of shaping, providing one of the most detailed models of animal shaping so far. Building upon and extending classical theories of both Pavlovian and operant conditioning (Rescorla and Wagner, 1972), they explicitly demonstrated well known effects such as fading and stimulus control in their model.

Agent shaping. In agent shaping the external environment remains constant over time. Instead internal properties or capabilities of the agent are changed resulting in easier learning problems. Amongst the internal parameters used for scaling are for example a reduced action set or complexity of state representation.

One such example of agent shaping is the work by Elman (1993). While the first part of his experiments simulated an explicit change of task, Elman was ultimately interested in automatic, intrinsic ways the network might achieve this staged learning itself. One key result was that similar effects to the improved learning through external staged learning could be achieved by resource limitations of the network itself. By deliberately limiting the working memory capacity of the network by degrading it over time with noise, only simpler sentences could be successfully processed and learnt. With slowly increasing capacity, the network learnt more and more complex sentences, but only after mastering the simpler ones. Thus, this structural non-stationarity of the network resembled the effects of shaping, leading to a form of "self-shaping". Elman argued that this change in working memory capacity could be accounted for by the developmental change of different parts of cortex and potentially explain why the slow maturation of (prefrontal) cortex is important for learning complex tasks like language.

Although the issue is not uncontroversial (see Rohde and Plaut, 1999), the idea that a form of self-shaping occurs through initial resource limitations in the network has gained some traction.

More recently, Reynolds and O'Reilly (2009) tested the idea of developmental self-shaping in their analysis of the emergence of hierarchical representations in a model of learning in prefrontal cortex (PFC) and basal ganglia (BG). Amongst several other manipulations of their network, that influenced the hierarchical nature of representations, they also analysed the effects of an initial reduction in working memory capacity, which expanded as the simpler elements of the tasks were learnt. They were interested to see if this had an effect of selectivity in representations with respect to the hierarchical nature of the task. Unlike Elman, they did not find a positive effect of shaping, at least on the particular metric of PFC representations they considered.

2.4.3 Learning to learn

The third facet of improved learning in steps is the concept of "learning to learn" (Thrun and Pratt, 1998). In this concept an agent typically trains upon a large number of tasks through which it can learn general properties or biases from the environment. As such, unlike the other two facets, learning to learn does not rely on any individual task in the sequence of tasks, rather the number of tasks observed determine the quality of the bias learnt.

Key to learning to learn is to learn to reduce or restrict the effective volume of search space in which subsequent learning occurs, reducing the amount of information necessary to find a good solution. A common method of achieving such reduced search or state space is a re-representation approach. In addition to the basic, full space, an agent can learn an abstract space grouping together states, or learn a reduced set of basis functions (Ferguson and Mahadevan, 2006). Alternatively, a softer approach of bias, is to learn a prior over restricted hypothesis spaces (Baxter, 1997).

Thrun and Mitchell (1995) considered two different scenarios of life long learning. In the simpler of these scenarios, all tasks in the "lifetime" of the robot took place in the same environment. The learning agent could therefore transfer knowledge about the environment itself. For instance, it could learn the state-action-state transition probabilities, i.e. a model of the environment. In addition to this model, an individual state value function could then be learnt per task. Thrun and Mitchel's second approach considered transferring knowledge between differing environments. Although transfer is much smaller in this setting, certain local invariants continue to hold across environments, especially as the robot itself, with all its sensors and effectors, remains constant. In both cases, Thrun and Mitchell (1995) were able to demonstrate that transfer of previously learnt knowledge is essential in scaling-up robot learning to more realistic scenarios.

However, there have also been other approaches like combining a learning to learn method with reward based shaping. Konidaris and Barto (2006), for example, have proposed an algorithm to learn from multiple related tasks a reward shaping function,

which could then be used in an auto-shaping fashion to speed up subsequent learning.

2.5 Gated working memory

The majority of the models presented in the previous section have their root in machine learning and robotics. As such, biological issues have played little to no role in their choice of modelling architecture. Although this thesis does not directly employ biologically detailed models either, it does aim to contribute to the conceptual understanding of natural learning and behaviour, and link between the two.

Out of the computational properties present in many of the abstract models, a number of key concepts emerge that are likely to play an important role in biological systems, too. Amongst these are at least: (1) the ability to store state and context, a function that can be sub-served by working memory; (2) a framework of rule representations to manipulate and update state representations in working memory, a property often attributed to prefrontal cortex; (3) a mechanism to protect working memory in light of distractors; (4) a process through which state representations effect actions, and (5) a process by which rules can be learnt from examples. On top of these basic requirements for an agent providing flexible behaviour, task sequential learning commands at least some additional ones: (1) a mechanism of modularisation to protect and facilitate a collection of competencies across task boundaries; (2) a mechanism to detect a change of task together with the ability to guide learning appropriately, and (3) a form of multi-tasking to integrate different tasks together again.

Particularly the first set of requirements, to model flexible behaviour, is well covered by the class of gated working memory models. They were therefore chosen as the basis of the current work, to inherit the systems level link to neuroscience, while equally providing a framework of computational abstraction for analysis. To show that the second set of requirements can also be well captured by these models is one of the main aims of this thesis.

This section therefore first briefly describe the basic properties of working memory in biological systems. It then reviews what is known about the neuroanatomical substrates of working memory and cognitive flexibility, the prefrontal cortex (PFC) and the basal ganglia. Finally, it discusses how all of this information is integrated to generate the class of gated working memory models.

2.5.1 Working memory

Working memory is the ability to hold an item of information transiently in mind, in the service of comprehension, thinking, and planning (Baddeley, 1986; D'Esposito, 2008). Particularly, it is the retention of information that is no longer accessible in

the environment, but necessary for guiding subsequent behaviour. Working memory encompasses both storage and processing functions, and is vital for nearly all flexible and cognitive behaviour. It is thought that the term was first introduced by Miller et al. (1960), although the concept of short term memory existed beforehand.

Activity based memory. Apart from its short time scale of decay, one of the key distinguishing properties of short term or working memory, in comparison to long term memory, is its method of storage. While long term memory is thought to be encoded in neuronal connectivity, working memory resides in persistent neuronal activity, a distinction that was already mentioned by Hebb (1949).

It is also widely believed that working memory is maintained actively in sustained firing of groups of neurons (e.g. Goldman-Rakic, 1987; Fuster, 1989). One of the potential mechanisms of this sustained activity is an attractor network (Hopfield, 1984; Zipser, 1991; Durstewitz et al., 2000b), i.e. a dynamical system with stable equilibrium states.

Capacity of working memory. Unlike long-term memory, short term or working memory is generally thought to have very strong capacity limitations of only a few items, possibly even as low as 4 (James, 1890; Miller, 1956; Luck and Vogel, 1997; Cowan, 2001). Nevertheless, the nature of this limitation is somewhat controversial. There are debates about the limit being hard or soft, item or time based, or storage or attention based. Furthermore, what is considered to be an object with respect to working memory capacity can vary and can even be trained for. This can sometimes complicate correct measurement of capacity and may underlie some of the discrepancies reported. This has led to the concept of "chunks", in which multiple objects can be combined into single chunks through training or experience.

Significant inter-subject differences in working memory capacity have been identified. One possible explanation is the different ability of individuals to filter or gate out irrelevant stimuli, and only represent task relevant information. For example, in an experiment in which subjects were asked to remember only a few select objects in a cluttered scene, Vogel et al. (2005) were able to show that general measures of working memory strongly correlated with subjects' ability to ignore the distractors in the display.

2.5.2 Neuroanatomical substrates of working memory I: The prefrontal cortex

The brain area most frequently associated with working memory and cognitive flexibility is the prefrontal cortex (Shallice, 1982; Baddeley, 1986; Fuster, 2008; Wood and Grafman, 2003; Miller and Cohen, 2001; D'Esposito et al., 2000). It has been implicated in nearly all aspects of cognitive flexibility, including goal directed behaviour (Balleine

and Dickinson, 1998), planning (Tanji and Hoshi, 2001), attention, set-shifting (Milner, 1963), prospective memory (Burgess et al., 2001) and working memory (Goldman-Rakic, 1995). There have been numerous electrophysiological studies showing persistent activity during delay time periods in various areas of PFC in non-human primates (e.g. Miller et al., 1996; Chafee and Goldman-Rakic, 1998; Quintana and Fuster, 1999; Funahashi et al., 1993, 1989; Fuster and Alexander, 1971). In humans, this finding has been supported by over 100 fMRI studies (e.g. Courtney et al., 1998; Zarahn et al., 1999; Sakai et al., 2002; Jha and McCarthy, 2000). Although one of the main foci of working memory appears to be in dorsolateral PFC (BA 46), different parts of the component model (Baddeley et al., 1974) of working memory are thought to reside in different areas of the PFC (Owen et al., 1999). For example, the phonological loop is mostly attributed to BA 40, both by lesion studies (Vallar and Papagno, 2002; Warrington et al., 1971) as well as fMRI (Paulesu et al., 1993). Other areas involved in working memory include inferior position in the prefrontal cortex (Wilson et al., 1993) and the dorsolateral PFC. Although in other areas outside the PFC small amount of persistent activation have also been reported (Fuster and Jervey, 1981), it is thought that only PFC has the ability to maintain it in the face of distractors (Miller et al., 1996).

Studies with anterior and retrograde tracing in areas 9 and 46 are beginning to dissect the local circuitry of PFC and working memory (Levitt et al., 1993; Kritzer and Goldman-Rakic, 1995; Pucak et al., 1996). It has been found that narrow, stripe-like bands of connectivity emerge. Furthermore, small clusters of neurons are connected. This stripe-like pattern is distinct from the more patch-like connectivity in posterior cortex. It has been suggested that this stripe-like organisation may allow for more localist, distinct or modular memory units (Frank et al., 2001a; Hazy et al., 2005; Goldman-Rakic, 1996), each independently encoding separate bits of information.

Other aspects of cognitive flexibility linked to the PFC. An additional important role for the PFC in flexible behaviour is maintaining representations of rules. Therefore, most tasks, that are disrupted following PFC damage, require the acquisition of conditional associations (if-then rules) (Passingham, 1993). Indeed, it has been shown that neuronal activity in PFC correlates with rule selectivity (e.g. Hoshi et al., 1998; Watanabe, 1990). For example, Asaad et al. (2000) trained monkeys to alternate between tasks that employed the same cues and responses, but had three different rules. They found that over half of active lateral PFC neurons have rule dependent firing. Similarly, Wallis et al. (2001) found that neurons in orbital PFC encoded whether the active rule was a "match" or "non-match" rule in a delayed (non) match to sample task.

Many of the roles and functions of PFC fall under the term of executive function (Gilbert and Burgess, 2008). At the heart of many of its theories is a distinction between routine (automatic) and non-routine (controlled) processing. It is thereby thought

that executive function exerts control over routine behaviour in order to suppress its actions when inappropriate (MacLeod, 1991). One of the potential ways of exerting such top-down control is the PFC's ability to maintain representations in order to bias lower levels of processing (Cohen et al., 2000). However to achieve its flexibility, PFC representations require a considerable amount of processing in the form of for example planning and hypothesis testing. To support such processes, an important ability is to be able to switch focus between stimulus oriented thought and stimulus independent thought, which has been suggested as one of the roles for anterior PFC (Burgess et al., 2007a).

However, PFC appears not only important in executing and representing rules, but also in learning (Eacott and Gaffan, 1992; Parker and Gaffan, 1998). One of the difficulties in rule learning is that it often involves learning associations between stimuli and behaviours that are separated in time. The persistent activity and working memory of PFC appears as an important substrate to bridge this gap.

Encoding of information in the PFC. Amongst the different theories of PFC function, a key aspect is whether they take a representational or processing view point (e.g. Wood and Grafman, 2003). In the representational framework, PFC holds information available in a way that more posterior circuits can use to augment current sensory information in their processing of actions, basically providing the necessary context to make the right decision. Examples of this approach are the working memory model of Goldman-Rakic (1987) or the guided activation theory of Miller and Cohen (2001). The processing approach on the other hand takes the view that cognition in the PFC can be described in terms of performance without specifying the representation that underlies these processes, and had been the traditional way of viewing PFC function.

Hierarchical organisation of the PFC. The prefrontal cortex is far from a single, uniform structure (Petrides and Pandya, 2002; Petrides, 2005). Although some of its separate areas perform quite distinct functions, there have recently been suggestions that there is a hierarchical organisation amongst parts of the PFC along a rostral-caudal axis (Koechlin et al., 2003; Badre and D'Esposito, 2009). Progressive rule abstraction is thought to occur along this hierarchy. The lowest level of abstraction includes simple associative rules specifying concrete stimuli and motor actions, which have been found to correlate with activity in pre-motor areas (PMd). One level up in the abstraction hierarchy, typically found in dorsolateral PFC, neurons can encode classes of rules. For example, a rule might indicate which modality should be used to select appropriate responses in a task. Finally, at the most anterior level, it is thought that the frontopolar cortex is mostly involved in "meta-level processing" (Koechlin and Hyafil, 2007; Burgess et al., 2007b).

2.5.3 Neuroanatomical substrates of working memory II: The basal ganglia

Neuroanatomy of the basal ganglia. The basal ganglia are a group of sub-cortical nuclei, consisting of the striatum, the internal and external parts of the globus pallidus (GPi / GPe), the substantia nigra pars reticulata (SNr), the substantia nigra pars compacta (SNc) and the subthalamic nucleus (STN). Afferent connections to the basal ganglia predominantly project to the striatum, making it the main input nucleus of the basal ganglia, although the STN also receives inputs (Mink and Thach, 1993). The output of the basal ganglia, predominantly to different parts of the thalamus, comes from the GPi and SNr nuclei. Although the two nuclei are anatomically separate, SNr and GPi are thought to be functionally and anatomically similar and are often considered as a single entity (Albin et al., 1989; Bolam et al., 2000).

One of the most striking anatomical features of the basal ganglia is their organisation into loops originating in cortex, projecting through the basal ganglia to the thalamus, and back to cortex. It is widely believed that the basal ganglia are split into 5 more or less distinct basal ganglia-thalamocortical pathways: "motor", "occulomotor", two "prefrontal" and a "limbic" loop (Houk et al., 1995b; Alexander et al., 1986; Alexander and Crutcher, 1990). Both the projections to the cortex (Alexander et al., 1986; Kemp and Powell, 1970) as well as the thalamocortical projections are organised in a topological fashion. In more recent years, the strict separation of the loops has been questioned, though, with both inputs and outputs converging from and to areas other than their main projection areas (Haber, 2003; Parthasarathy et al., 1992).

Another striking feature of the basal ganglia is their convergent nature of the pathways. Nearly all of the cortex projects to the striatum (Carman et al., 1963; Kemp and Powell, 1970), targeting a much smaller area, that itself projects down to the even smaller nuclei of the SNr and GPi. For example in humans, the full cortex projects down to a striatum that is thought to have on the order of 110 million neurons (Fox and Rafols, 1976). These further converge on only about 160,000 neurons in each the GPi and SNr (Lange et al., 1976). The loops have often been described as a form of "funnel" (Allen and Tsukahara, 1974). It has therefore been suggested that this convergent nature makes the basal ganglia well suited for generalisation and dimensionality reduction (Bar-Gad et al., 2003).

Pathways of the basal ganglia. The basal ganglia are thought to have two main pathways from the input nucleus of the striatum to the output of the GPi and SNr, the direct and indirect pathways. In the direct pathway, the striatum projects to the GPi and SNr directly, while in the indirect pathway striatum projects to the STN through the GPe and then on to the GPi. The output of the basal ganglia is tonically active and inhibitory in nature, suppressing input from the thalamus to cortex. As output

from striatum is also of inhibitory nature, firing in the direct pathway acts to disinhibit thalamus. In contrast, the additional inhibitory synapse from GPe to STN results in the indirect pathway disinhibiting the GPi, and inhibiting the thalamus. Due to the opposite effects on thalamus, these direct and indirect pathways have also been called the "Go" and "No-Go" pathways.

Role of the basal ganglia in flexible behaviour and working memory. The basal ganglia are also thought to play an important role in cognition. For example in studies by Goldman and Rosvold (1972), monkeys with striatal lesions showed similar impairments on a typical prefrontal test battery as monkeys with lesions to the dorsolateral PFC. Similarly, patients with Parkinson's or Huntington's disease, disorders known to primarily affect the basal ganglia, are impaired in sorting or planning tasks, use of memory stores, and manipulation of internal representation (Dubois and Pillon, 1996; Gotham et al., 1988; Taylor and Saintcyr, 1995; Brandt and Butters, 1996). Furthermore in healthy subjects, set-shifting has been linked to increased activation of striatum (Rogers et al., 2000).

2.5.4 Role of the basal ganglia in learning

One of the most extensively studied functions of the basal ganglia is their involvement in stimulus - response style habit learning (Packard and Knowlton, 2002). This has been shown in a variety of different studies covering at least rats, monkeys and humans (e.g. Graybiel, 1998; McDonald and White, 1993; Packard et al., 1989; Teng et al., 2000; Butters et al., 1994; Knowlton et al., 1996; Martone et al., 1984). Several of these studies have contrasted the S-R learning of basal ganglia with the contextual or spatial learning of hippocampus, dissociating the two in a number of lesion studies.

In contrast to the flexible and rapid updating behaviour of goal-directed learning in PFC, S-R learning in the basal ganglia is thought to be slower and independent of current goals (Yin and Knowlton, 2006). One group of experiments nicely demonstrating this are the reinforcer devaluation studies (Dickinson and Balleine, 2002). In these experiments, animals are trained on a task such as lever-pressing for which they receive a reward. After extensive training, the reward is devalued for the animal, e.g. by injection of the aversive drug lithium chloride, which makes the animal associate the food reward with sickness. In goal-directed lever-pressing, this abolishes lever pressing in a subsequent test, since the animal no longer has a desire to obtain the reward. In case of habit formation, however, the animal continues to perform the trained action.

This stimulus-response learning appears to correspond closely to parts of reinforcement learning. With their strong innervation by dopamine, which has been linked to temporal difference (TD) learning (Schultz et al., 1997; Niv et al., 2005), as well as their convergent nature of connectivity, the basal ganglia are well positioned to play the role

of the action selection network in the reinforcement model (Joel et al., 2004). Furthermore, many fMRI studies have shown the involvement of the basal ganglia in reward learning (Delgado et al., 2000; Knutson et al., 2000; Pagnoni et al., 2002; O'Doherty et al., 2003; McClure et al., 2003, 2004; Rodriguez et al., 2006), strengthening the hypothesis of reinforcement learning in the basal ganglia.

Although emphasis of learning in the basal ganglia has predominantly focused on habitual style learning, the strong involvement of the basal ganglia in more flexible goal-directed behaviour suggests that the basal ganglia might play an important role in learning flexible behaviour, too. For example, the basal ganglia have been implicated in sequence and categorisation learning (Seger, 2006). However, its picture is less clear so far, but experiments and models, such as the the gated working memory models, are increasingly beginning to dissect and probe this aspect of learning as well.

Role of the basal ganglia in gating. One hypothesis of the effect of the basal ganglia cortical loops on motor control is to gate actions (Chevalier and Deniau, 1990; Schneider, 1987). Rather than affecting motor cortex directly, it enables or prevents an action in response to stimuli, thus "opening the gate". For example in experiments silencing the output of the basal ganglia, reactivity of cortical neurons to sensory inputs increased. (Chevalier et al., 1985; Buee et al., 1986)

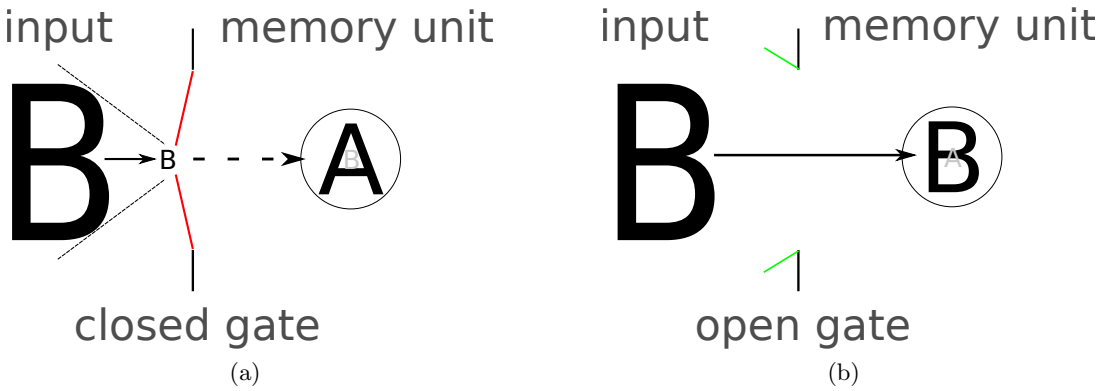
2.5.5 Models of gated working memory

Combining various of the above components, the class of gated working memory has emerged as a powerful set of models for capturing flexible behaviour.

Although attractor networks can be constructed to be arbitrarily stable against noise, this comes at the expense of making it increasingly hard to update them, and thus to store new bits of information when needed. However, one of the important features of working memory is its ability to update rapidly. Furthermore in the domain of cognitive control, distractors may often elicit equal, or even stronger, responses than the task relevant stimuli themselves. Thus, inherent to working memory is the dilemma between stability and plasticity (Grossberg, 1980; Carpenter and Grossberg, 1988; Bengio et al., 1993; Mozer, 1992).

One possible solution to this is the use of active gating of memory, models of which date at least back to Zipser (1991); Servan-Schreiber and Cohen (1992); Cohen et al. (1996). It has since gained substantial traction in many different models (e.g Cohen et al., 1996; Hochreiter and Schmidhuber, 1997; Beiser and Houk, 1998; O'Reilly et al., 1999; O'Reilly and Munakata, 2000; O'Reilly and Frank, 2005; Rougier et al., 2005; Dayan, 2007; Todd et al., 2009).

The idea is to have a dynamic gating signal that acts as a multiplicative gain control



(a) (b)

Figure 2.1: Cartoon of a gated working memory: (a) If the gate is closed, the amplitude of the inputs are modulated down to not be strong enough to overcome the predominant memory and replace it. (b) When the gate is open though, inputs are passed through to the memory in full strength. If the input weights for the presented pattern are large enough, the inputs can then be stored in the memory.

on the inputs to memory. Importantly, this signal is under the learnt control of the model itself. Its multiplicative component allows it to overcome this trade-off (depicted in figure 2.1). With a closed gate, the low multiplier tunes down inputs relative to the strength of the memory, resulting in inputs being not sufficiently strong to overcome the attractor dynamics. An open gate, however, allows the full strength of the input connections into memory and hence to rapidly update its current state. There have been several suggestions as to how such a proposed gating mechanism may be implemented.

Stabilising attractors with dopamine. One prominent hypothesis is that dopamine itself can directly drive gating to influence the strength of the attractor (Servan-Schreiber and Cohen, 1992; Braver and Cohen, 2000) and therefore the ease of updating. It is based on experimental evidence of the importance of dopamine signals for working memory (Luciana et al., 1992; Sawaguchi and Goldman-Rakic, 1991; Williams and Goldman-Rakic, 1995; Penit-Soria et al., 1987), their appropriate timing (Schultz et al., 1993; Watanabe et al., 1997) as well as biophysical simulations, showing the ability of dopamine to effect the stability of delay period activity (Durstewitz et al., 2000a; Brunel and Wang, 2001). This hypothesis has been successful in modelling some often studied cognitive control tasks.

However, the dopamine signal is relatively global, spanning large areas of PFC in a potentially undifferentiated manner. In complex tasks, selective gating of various different elements of memory at different times is required. It is therefore difficult to imagine this signal to be sole responsible for gating directly (Dreher et al., 2002; Frank et al., 2001a).

Gating working memory with the basal ganglia. An alternative hypothesis is that the gate to working memory is driven by the executive loop of the basal ganglia

- thalamo - cortical system, using the topologically refined projections of striatum to cortex. It is known from various studies (e.g. Brown and Marsden, 1990; Harrington and Haaland, 1991; Willingham and Koroshetz, 1993, for more details see 2.5.3) that the basal ganglia are involved in cognitive control functions. Further, the similar role of the direct and indirect pathways in gating action selection in motor control (e.g. Bullock and Grossber, 1988; Neafsey et al., 1978; Schneider, 1987; Mink, 1996) make the executive loop of the basal ganglia a prime candidate to suggest for a role in gating working memory (Chevalier and Deniau, 1990). Finally, the high convergence ratio from cortical projections (Kincaid et al., 1998), the competition between neurons through collateral inhibition (Plenz, 2003) and the strong innervation of dopamine neurons relevant for learning (Houk et al., 1995a), equip the basal ganglia with the necessary pre-requisites to classify patterns in order to decide when to gate.

One of the first computational models to use the basal ganglia - thalamo - cortical loops as a form of gated working memory was presented by Dominey and Arbib (1992). Their model simulated the production of spatially accurate sequential saccades. The model contained many components to handle the motor output and sensory input mapping. Importantly, it also included recurrent connections in form of loops between the frontal eye field and the thalamus, to act as working memory to store the position of the saccade sequence. Each loop, one per possible spatial position of a saccade, was inhibited by tonically active model neurons from the SNr, which in turn included inhibitory connections from neurons in the caudate nucleus. Only when thalamic neurons were disinhibited could the recurrent activation between frontal eye fields and thalamus be initiated and continue once the inhibition was restored. Thus, they acted as a gate. Unlike many later models though, caudate neurons were not modelled using convergent connections in order to drive decision of when to gate.

Shortly after, Houk and Wise (1995) presented a conceptual model of the interaction between the basal ganglia and cortex. They postulated a basic set of building blocks that could form a distributed model for the control of action, linking it to a wealth of anatomical detail of the basal ganglia and PFC. This conceptual model was later implemented by Beiser and Houk (1998) in a network of dynamical rate based neurons to simulate the encoding of serial order of sensory events. In the process, much of the anatomical detail was simplified or removed. Critical to what remained, however, was a collection of memory modules, formed by loops between PFC neurons and thalamus. Here too, the disinhibition of thalamus through basal ganglia interactions was needed to engage working memory. However, activation patterns of memory, being akin to binary units, were not directly dependent on other inputs. Strictly speaking, they did not fulfil the definition of a gate, though many of its aspects remained. Furthermore, the model's success in encoding sequential patterns came from the selective ability of caudate to trigger working memory, based on recognising diverse neuronal patterns in other parts of the model.

As both approaches to gating are compelling, Gruber et al. (2006) aimed to reconcile both approaches into a single model. While dopamine in the PFC enhanced global robustness of attractors, guarding it against external noise or disruption, the focused projections of the basal ganglia enhanced robustness to internal noise by deepened individual attractor states.

Sequential learning of when to gate. The above models of basal ganglia gating neglect the question of learning when to gate, and relied either on random weight connections or on hand tuning of parameters. O'Reilly and Frank (2005) introduced more recently a model combining the anatomical ideas of the previous accounts with computational ideas of learning in gated memory. The "Prefrontal, Basal ganglia Working Memory model" (PBWM), as they termed it, combined ideas from a machine learning model of gated memory (LSTM, Long Short-Term Memory) (Hochreiter and Schmidhuber, 1997; Gers et al., 2000, 2003), with the computational role of dopamine in reinforcement learning (Schultz et al., 1993, 1997; Dayan and Niv, 2008), to form a powerful model of cognitive control.

Chapter 3

Models and methods

Later chapters directly depend on a variety of models of gated working memory. Specific implementations of these models as they are used in simulation are described, including the LSTM used in simulations in chapter 4 and the bi-linear model of Dayan (2007). Although PBWM was not directly used in any of the simulations, it was both influential to current working as well as the basis of several of the descried models, warranting a detailed review.

3.1 Prefrontal Cortex, Basal Ganglia Working Memory model

The Prefrontal cortex Basal ganglia Working Memory model (PBWM) (Frank et al., 2001a; Hazy et al., 2005, 2007) is an abstract modular neural network architecture. Its modules were designed to follow some of the basic structural and functional principles known from systems neuroscience. The model consists of several modules, each representing an area of the brain such as "the PFC", the "posterior cortex", "striatum" or "substantia nigra". The modules consist of standard neurons augmented with area-specific functional specialisations. Later variants also included a combined reinforcement, supervised and unsupervised learning system. With the help of this model, Frank and colleagues explained a number of theoretical concepts. They have made behavioural predictions, several of which have been experimentally demonstrated in both healthy and Parkinson's subjects.

The basic unit of a neuron in this model is derived from the Leabra framework ("learning in an error-driven and associative, biologically realistic algorithm") (O'Reilly, 1998; O'Reilly and Munakata, 2000; O'Reilly, 2001). Rate based activation of these point-neurons are governed by a conductance-based dynamical system of their membrane potential. Conductances come in three forms: The excitatory or input-driven conductance realises the standard neural network ideas of a weighted average of activations of

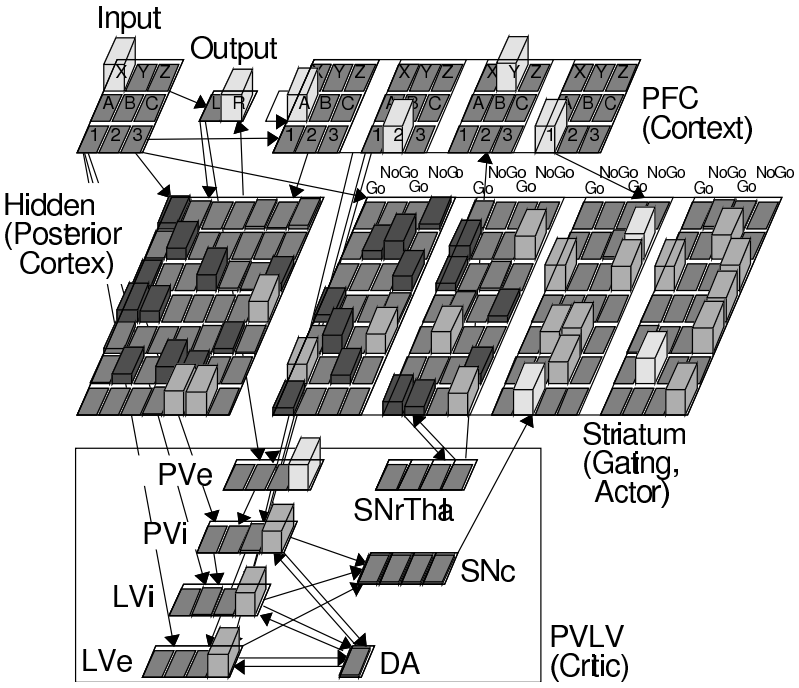


Figure 3.1: Prefrontal cortex, Basal ganglia Working Memory model: This model consists of 5 basic components. The input / output layer, a standard hidden layer termed the posterior cortex, a special memory layer termed PFC, the gating layer and a reinforcement learning component. The hidden layer draws inputs from the input/output layer and the memory layer. It allows for a more expressive representation due to its converging expansion. For the memory layer, whenever a block's (stripe's) associated gate is open, the activation of the input layer is copied and stored in the respective memory. Figure taken from O'Reilly and Frank (2005).

its inputs. The inhibitory conductance realises a "k-winner takes all" component and is uniform across all units of a module. It is chosen to suppress firing in all units, bare the k most active ones. The final conductance is a simple leak current.

Learning of the synaptic weights is driven both by an unsupervised / Hebbian component similar to the Oja rule (Oja, 1982), as well as a type of supervised learning akin to contrastive Hebbian learning (Hinton, 1989; Movellan, 1990).

The architecture of the network, shown in figure 3.1, consists of 5 basic components and an RL component. The first of these is the input / output layer, and its representations are therefore determined by the environment and the task modelled. The second component, the "posterior cortex", is a standard hidden layer in the network. It acts as an integrative layer between all other components, drawing inputs from the input / output layer and the working memory layer. It projects to the majority of components of the network.

The third component, the memory layer or "PFC", is further subdivided into multiple components. Each of these components is a single memory module, or as it is named in this model "a stripe", in reference to anatomical organisation of PFC described in Levitt et al. (1993, 1996). Whereas in earlier versions, PFC representations were

direct copies of the input layer, Reynolds and O'Reilly (2009) managed to use fully learnt connections. Unique to this layer is its intrinsic bi-stability under the control of the corresponding go / no-go firing of the striatal neurons. This endowed it with the properties of a gated working memory.

The most interesting and complex layer is that of the striatal gating units. Subdivided yet one level further, each striatal stripe consists of a set of "go" and "no-go" neurons. In addition to standard neural properties, activity is driven by the dopamine levels, expressed by SNc firing, enhancing go neurons and suppressing no-go neurons. Summed-up and contrasted in the SNr/thalamus layer, these neurons trigger bi-stability in PFC neurons. They cause gating as well as multiplicatively modulate the per-stripe dopamine neurons, enhancing its effect at the time a gating operation occurred.

Dopamine implements a form of reinforcement learning. However, unlike the temporal difference (TD) interpretation of dopamine found in many other models, PBWM introduces a new learning mechanism called PVLV (primary value, learnt value) (O'Reilly et al., 2007). It explicitly distinguishes itself from the TD interpretation, stating that the sequential chaining of predictions required for learning with TD is not applicable to the typical tasks, PBWM was designed for. The PBWM also argued that a simpler, Rescorla-Wagner rule (Rescorla and Wagner, 1972) derived system was more biological plausible. However, Todd et al. (2009) showed that with the help of eligibility-traces some of these limitations could be overcome to successfully learn.

This model has had impressive successes in explaining a wide variety of cognitive and behaviour results. However, perhaps inevitably for a system designed to be somewhat realistic, this comes at a price of resulting in a highly complex and varied system. The inclusion of three different forms of learning (unsupervised / Hebbian, supervised learning and reinforcement learning) makes it particularly difficult to distinguish the individual contributions of each component. The simulations of this thesis don't use this integrated model of gated working memory, instead they explore the various aspects with simpler models, each of which only employs a single method of learning.

3.2 Long Short-Term Memory model

The computational principles of gated working memory embedded in PBWM stem from the Long Short-Term memory model (LSTM) (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) LSTM derives from a history of formal neural network theory in the field of machine learning, focused on computational aspects of learning sequential tasks with long time lags. Nevertheless, the LSTM was influential on neurological models of gated working memory, particularly on PBWM by providing the computational foundations of learning in gated activity based memory models. Particularly, it provided the first demonstration of how powerful learning in gated architectures could be. LSTM has

been successful in being used for solving a diverse set of problems, such as time series predictions (Gers et al., 2001), finding temporal structure in music (Eck and Schmidhuber, 2002), handwritten recognition (Graves and Schmidhuber, 2009) and speech recognition (Graves et al., 2004).

The context for LSTM is that recurrent neural networks are extremely powerful and capable of performing any sequential task (Schäfer and Zimmermann, 2006), including cognitive tasks such as those discussed in this thesis. However, training these networks was long believed to be very hard, if not close to impossible for long-term dependencies (Bengio et al., 1993). For example, one of the most commonly used training algorithms for recurrent neural networks, "Back-Propagation Through Time" (BPTT) (e.g. Williams and Zipser, 1992, 1995; Werbos, 1988) is a generalisation of the supervised learning gradient based algorithm back-propagation (Rumelhart et al., 1986). In short, BPTT acts in a way akin to unrolling the recurrent network, making it a deep feed-forward network, with one layer per time step. As with standard back-propagation, weights can be updated to minimise the averaged squared error of the output activations by descending along the gradient. But although BPTT works rather well for shallow networks and thus for short temporal sequences, the gradient signal tends to either numerically grow without bound, or rapidly diminish the further it has to be propagated (Bengio et al., 1994). Indeed, as analysed in Hochreiter (1991), the multiplicative nature of the chain rule of differentiation makes the error signal scale exponentially with the distance it has to be propagated. BPTT therefore cannot learn over long time delays. Other training algorithms, such as for example Real-Time Recurrent learning (RTRL) (Robinson and Fallside, 1987) or Time-Delay Neural Networks (Lang et al., 1990), as well as combinations thereof (Schmidhuber, 1992), suffer similar exponential scaling.

LSTM was designed to overcome the exponential scaling, even for input/output relations spanning hundreds of time steps. Its solution is to enforce a self-connection with a fixed weight of one, creating a memory unit. This way, together with the inclusion of multiplicative gates to protect the contents of the input against distractors, it captures much of the essence of the more biologically derived models of gated working memory. LSTM also has a multiplicative output gate to suppress the influence of its memory on the environment until it is needed. Indeed, Gers et al. (2000) introduced a further gate, the "forget gate", allowing the network to forget or clear out a memory when no longer needed. This is a necessity to handle continuous environments where trials overlap and no natural point of "reset" exists. In later variants (Gers et al., 2003), the model was further extended to include ungated, direct connections from the memory cells back to the gating layer. This was necessary as in the standard version, the content of cells whose output gates were closed could not influence the decision of its own gating routines. However in this thesis, only the LSTM variant with forget gates will be considered.

As the LSTM network underlies all simulations in chapter 4 and some in chapter 6, it is now described in detail:

3.2.1 The network model

The LSTM network can be decomposed into three layers: input, recurrent or memory, and output layer (from bottom to top in figure 3.2). The output layer is a standard connectionist neural network layer, i.e. a linearly weighted sum of inputs with a sigmoidal activation function. The activations of the input layer are set to represent the percepts of the network and are represented as a unary encoding of the stimulus alphabet, although other encodings are equally possible. These are fed forward to both, the recurrent memory layer and directly to the output layer.

The recurrent layer consists of several so called memory blocks. Each memory block (the inset in figure 3.2) is associated with a set of three gating units, input, forget and output gates. In addition, each memory block has several memory cells, which store the actual memory through their recurrent self-connection (depicted by the middle circle of the cell in figure 3.2). Their activation can range from positive to negative real values. In the present simulations, these were bounded between -20 and 20 to prevent numeric overflow in the output non-linearity. Since multiple memory cells share a common set of gates, a more sophisticated encoding within a memory block can emerge, either allowing for a higher information content per block or representing it in a way easier to use in later processing stages. At each time step, the activation of a memory cell decays multiplicatively according to the activation value of the forget gate of the corresponding memory block. Thus, the forget gate interpolates smoothly between forgetting the memory completely (value of 0) and retaining perfect memory (value of 1). Similarly, the input to the memory cell, a linear combination of the input layer and the outputs of all of the memory cells passed through a sigmoidal non-linearity ranging from -1 to 1, is multiplied by the activation value of the input gate before being added to the activation of the memory cell. This allows inputs to be differentially ignored. Finally, the output of a memory cell is passed through another sigmoidal non-linearity, and is multiplied by the activation of the output gate unit.

All sets of weights, i.e. for the output, cell input and gating units (depicted by triangles in the figure), are fully plastic. This allows the network to learn to select between rapid updating and robust memory retention, wherever required. Note that the LSTM contains no direct competitive component on the output layer. Any anti-correlation between its units has to be explicitly learnt through the weight vectors of the output units. Potential failures have to be dealt within the mapping from output units to decisions of the network, where necessary.

The output units, as with all other units in the network, have real valued activations. For environments in which the output is discrete, such as those presented in chapter 4,

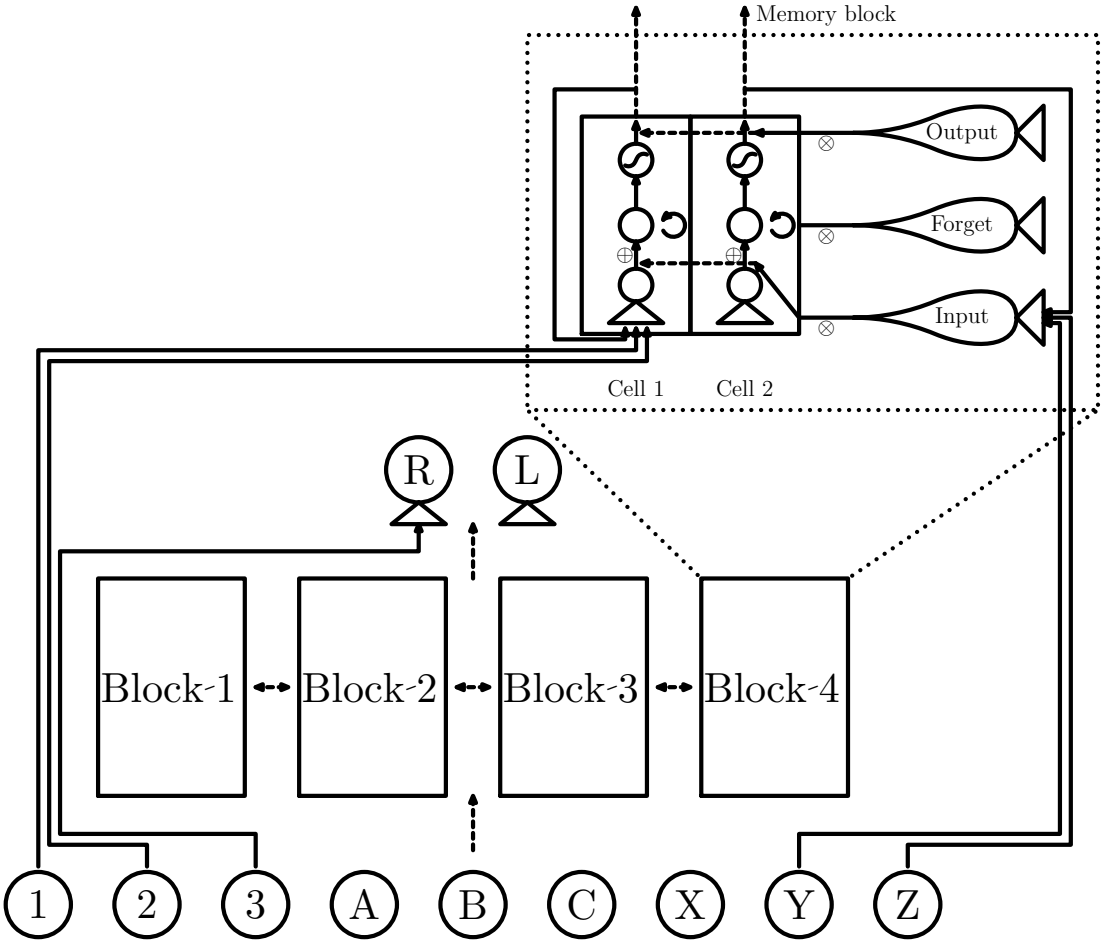


Figure 3.2: The LSTM model: The Long Short-Term Memory (LSTM) network with 'forget' gates. Stimuli are encoded punctately in the input layer of the network using binary units that each represent one element of the alphabet. The number of input units used varies between 9 (for the main `12-AX` case) and 18 (for both shaping and reversal tasks). All networks have two output units representing an 'L' or 'R' decision of the network. The hidden layer comprises between 4 and 8 recurrently connected memory modules or blocks. Feed-forward weights connect the layers. A close-up of a memory block is shown, visualising the two cells commonly gated by multiplicative in, out and forget gates. Each triangle represents a set of linear weights receiving inputs from the input layer and the gated output of all the memory cells.

the graded activations of the output units have to be binarized. In the present simulations, units were thresholded at an activation of 0.5, though the error signal for supervised learning was computed on the graded values.

The detailed equations used in this thesis in chapter 4 and chapter 6 can be found in appendix A.4.1 or in the original paper describing LSTM with forget gates by Gers et al. (2000).

3.2.2 Learning in the LSTM model

Learning in the LSTM network combines a variant of back propagation through time (BPTT) (Williams and Peng, 1990) with a modified form of the real time recurrent learning algorithm (RTRL). The self-connections of the memory cells, i.e. the constant error carousel, form a boundary, splitting learning into the two different algorithms. Units that are on the output side of the memory cells, i.e. the output units themselves as well as the output-gating units, are trained with the BPTT variant. Due to the recurrent properties of the network, activity of the output-gates and the resulting error gradients do feed back into the memory units and therefore are both on the input and output side of memory. However for simplicity and computational efficiency, these gradients are truncated, such that the error is assumed to only flow directly through the memory layer. All weights that are on the "input side" of the memory layer, i.e. the weights into the cells, the weights to the input gates and the weights to the forget gates, are trained with the help of the RTRL algorithm instead. Here too, all gradients not flowing through the memory unit are truncated. By nature of this truncation process, learning in the LSTM network with forget gates remains local both in space and time and leads to efficient learning of all weight connections.

3.3 A RL actor-critic model of gated working memory

A further, abstract, model of gated working memory is the actor - critic based reinforcement learning model presented by Todd et al. (2009). This model uses standard RL algorithms like TD(λ) on top of an actor - critic framework, unlike the PBWM model, which uses its own specialised RL algorithm PVLV (O'Reilly et al., 2007) as one of three learning systems. It also lacks much of the complexity and sophistication of the PBWM, instead concentrating on a pure and theoretically sound model. Nevertheless, it can be successfully applied to the same type of tasks as the PBWM, showing that temporal difference based models are in fact capable of learning tasks providing only weak temporal chaining due to their stochasticity. This counters the suggestions of O'Reilly and Frank (2005).

Key to successful learning is the use of eligibility traces. These can be interpreted in terms of TD(λ), a method for smoothly interpolating between the one step backup

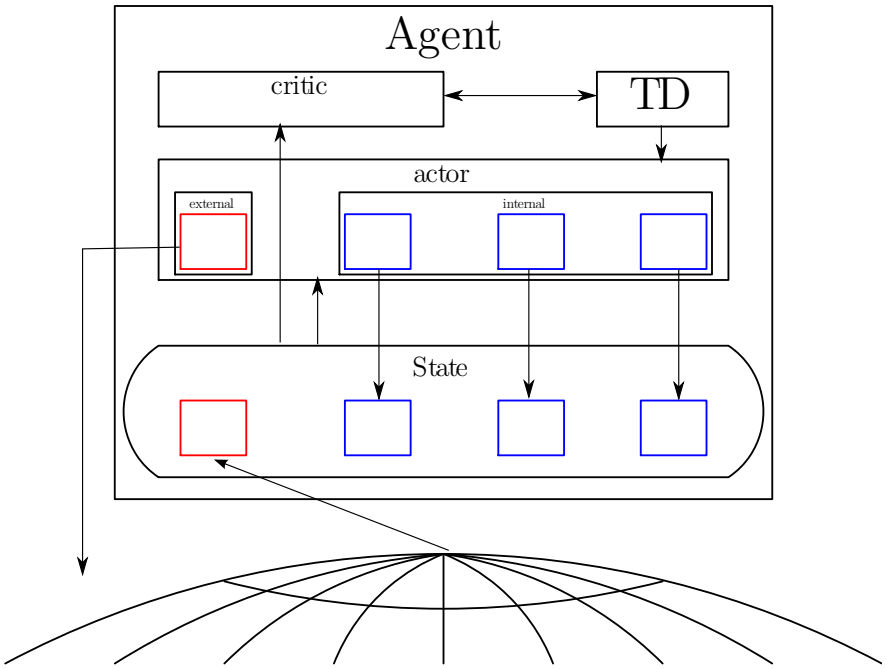


Figure 3.3: A reinforcement learning based model of gated working memory: The state space upon which the model acts consists of the primary percept coming from the environment, as well as a set of working memory modules which store previous percepts. These memory modules are subject to the gating actions of the internal actors, of which there is one per memory module and through which the model can update and learn its own state space representation. The external actor, together with the critic, form a standard actor-critic model which is trained through a temporal difference learning algorithm with eligibility traces.

of TD(0), and Monte Carlo sampling, working on the full sequence. Eligibility traces provide an infinite, albeit decaying, memory of past states beyond the explicitly gated memory of the model. This allows it to overcome issues of weak chaining until the correct gating policy has been learnt.

Todd et al. (2009) provided a theoretical account of the underlying computational principles for much of the class of gated working memory. Framing it in a way of solving problems in the domain of partially observable Markov models (POMDP) allowed to gain insights into the wider group of tasks, for which these models are suitable. This describes the purpose of gated working memory to explicitly support optimal behaviour (in terms of discounted future reward) in a POMDP. An important aspect of this solution is the model's ability to influence and learn its own state space representation, given by the inclusion of controlled memory elements into the state vector.

The model is an extension of the classic actor - critic architecture (Sutton and Barto, 1998), which in the past has been implicated as a model of basal ganglia operation (Joel et al., 2004), to the realm of the class of gated working memory models. The main difference to the standard actor - critic lies in the definition of the state vector. In addition to the usual immediate observation (current input), the state vector includes several controlled memory elements, equivalent to the working memory modules of other models. As such, the state vector is no longer uniquely defined by the outside world alone. Instead, it is also controlled through the model's own actions. These determine how the true world state of full stimulus history maps onto the internal state representation, driving the agents actor and critic.

In order to control these elements, the actor network is extended to include several further independent internal actors on top of the standard external or motor actor, one for each memory element. The internal actors act to store and memorise the current observation in their respective memory elements until the next gating action. Thus instead of one actor, the model is comprised from N+1 independent and different actor networks.

The critic network provides an estimate of the expected discounted future reward. It is unmodified from the standard actor - critic setting, albeit its operations on the non-stationary augmented memory state space.

3.4 A Bi-linear model of rules and habits

The bi-linear gated working memory model introduced by Dayan (2007) is more abstract than several of those described earlier, and not directly concerned with biological plausibility. However, unlike the LSTM network, which was devised to overcome some specific machine learning problems, Dayan's aims were to illustrate several important concepts of cognitive behaviour and its links to the structure believed to sub-serve

these, the prefrontal cortex, the basal ganglia and episodic memory.

Of particular interest was the ability, known at least for linguistically capable humans, to perform a new, complex task instantly by instruction, without the need for extensive training. This is another aspect of human flexibility that has so far mostly been neglected in the study of neural modelling. Although language itself is likely to play an important role in this instantaneous adaptation, there seems to be an additional translation process converting episodically recalled verbal instructions into a weight and activation based encoding in prefrontal cortex. This process may be seen akin to the ideas of compiling a programming language text into machine executable format, although this process may be rather more lossy in the case of natural agents like humans (Duncan et al., 2008). The substrate in which such 'compiled' instructions may be executed, and its relation to the more typically learnt patterns through trial and error, were the focus of his model.

Since it focused on instant learnability and instruction, the model introduced two important components beyond the classical group of gated working memory models: A concept of rules and a rule storage / retrieval mechanism. The idea of rules was to break the full task into several very simple rules, like "if you have seen an X then press the right button". These could conceivably be verbally instructed, although that part was not specifically modelled.

Key to the overall architecture, though, was that rules in the end used the same computational substrate of bi-linearity as the habitually learnt tasks, the main difference being the level of complexity in each. Although this way, tasks require several rules, each rule forms a more sparse representation.

3.4.1 The network model

As described above, the model comprised two main components, rule (or habit) execution, which determined the output of the model and manipulated the internal state and the rule retrieval and matching.

As a member of the family of gated working memory models, the state upon which the model selects its current actions consists of both the current perceived state of the world as well as a set of memory units storing a select set of previous percepts. The action of the network further consists of internal actions specifying when and what to gate into these memory units, as well as the actual action in the external environment. In addition, as each task consists of a complete rule set of several rules, a further internal action determines, if a given rule should in fact take any effect, or if it is currently not applicable. Thus, the model Dayan (2007) simulated consisted of 6 output units.

Each rule drives all of the output units o^c through a bi-linear function, whose activation is chosen stochastically according to

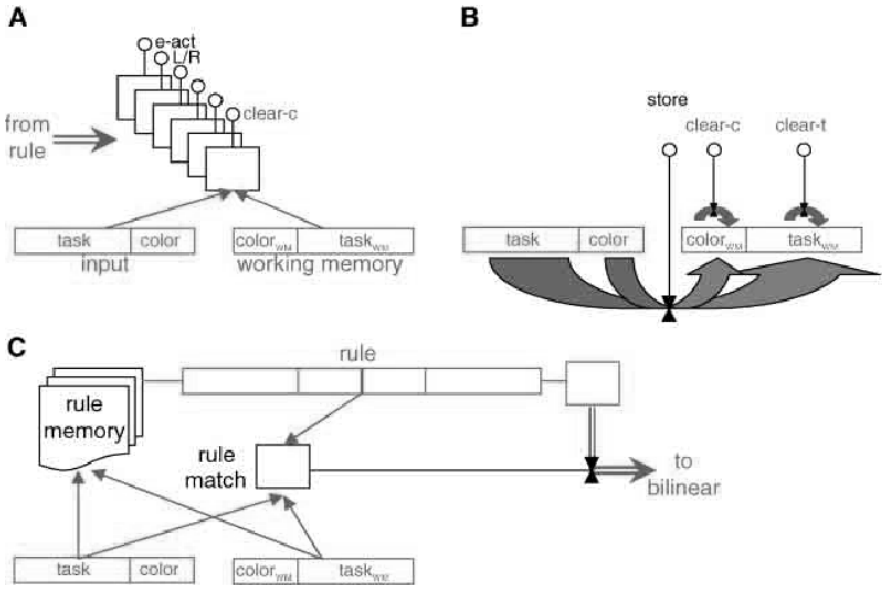


Figure 3.4: The Bi-linear gated working memory model: A) The rule execution structure: Each rule is split into several binary output units who each represent one of the available non exclusive actions. An action can either be external and act upon the world, or internal and manipulate the state of working memory. Working memory is thereby split into distinct memory units, or stripes (In this example into two for task and colour memory). B) Internal actions gate inputs into working memory and allow to selectively forget or clear memory. C) A task consists of a collection of rules in rule memory that can be retrieved in an associative step and and matched against for execution. Figure taken from Dayan (2007).

$$P(o^c = 1) = \sigma\left(\sum_{ij} x_i W^c_{ij} x_j + \sum_i u^c_i x_i + b^c\right) \tag{3.1}$$

The state vector $\vec{x}$ thereby is a simple concatenation of the binary representation vectors of the current percept with the vectors of each of the memory cells of the model. σ is the standard logistic function $\sigma(\xi) = 1/(1 + \exp(-\xi))$

However, the model assumes that there may be a rather large set of rules, an agent may be capable of performing overall. As such, it is unlikely to be feasible to execute each one of the rules at every given time point, to determine the correct set of internal and external actions. Therefore prior to execution, the rules need to be retrieved from an episodic, auto-associative "storage and recall device". This role is expected to be fulfilled by the hippocampus and is depicted in figure 3.4c. This recall was modelled as an associative "matching" between elements of the rule and the current internal state of the agent. Unlike the more precise bi-linear execution action of the rule itself, the associative episodic matching, however, does not provide the ability of exclusion criteria, i.e. specify conditions under which it won't apply. As such, this process may retrieve multiple rules, all of whose weights are in turn instantiated and executed in the fashion described above.

In contrast to the original bi-linear model described here, the model used in chapter 6 included a few modifications to the way internal actions are represented and the way an active rule from the rule set is chosen. These differences are described in more detail in the chapter itself.

3.4.2 Training the model

The bi-linear model of rules and habits takes a different approach to learning than the previous models of LSTM and the PBWM. As its focus was on the computational substrate of a rule based mechanism, potentially capable of instant reprogrammability in the form of verbal instruction, it did not asses trial and error learning. Instead a more direct supervised learning approach was employed, assuming a form of teacher-forcing (Williams and Zipser, 1989), to avoid the need for temporal credit assignment. Each of the rules was trained separately using standard gradient decent maximum likelihood. A large set of training examples was generated to train the bi-linear weights, each providing the mapping from a given internal state representation to the correct output.

3.5 The 12-AX task: Testing the benefits of gated working memory in learning

There are already a substantial number of tasks designed to test and study working memory, flexible behaviour and its neuronal substrates (Burgess, 2010). These span a wide range of complexities from simple laboratory tasks, like the n-back working memory task or Stroop task which focus on specific aspects, to open ended, complex, every day living tasks like the Multiple Errands test (Shallice and Burgess, 1991) which draw on a wide range of skills. This thesis aims to study the computational properties of task sequential learning and the architectural support required to benefit from complex sequential information provided by the sequential breakout into sub-tasks. For this purpose a number of specific properties are necessary in a task beyond those found in more simple behavioural tasks. These include:

- Hierarchical decomposition: Absolutely key to studying task sequential learning is the ability to break the full task into a set of sub-tasks that each in their own right are simpler to learn, i.e. require a lesser amount of time to master in a trial and error learning style. For task sequential learning, the final task could be a simple concatenation of its individual components. However, more interesting is the case where the top level task contains certain components of the sub-tasks, but is still an integral task in its own right, i.e. is more than just its parts. Under these conditions, the information from the sub-tasks can be used to benefit from learning in stages, but this information has to be adapted to meet the new requirements

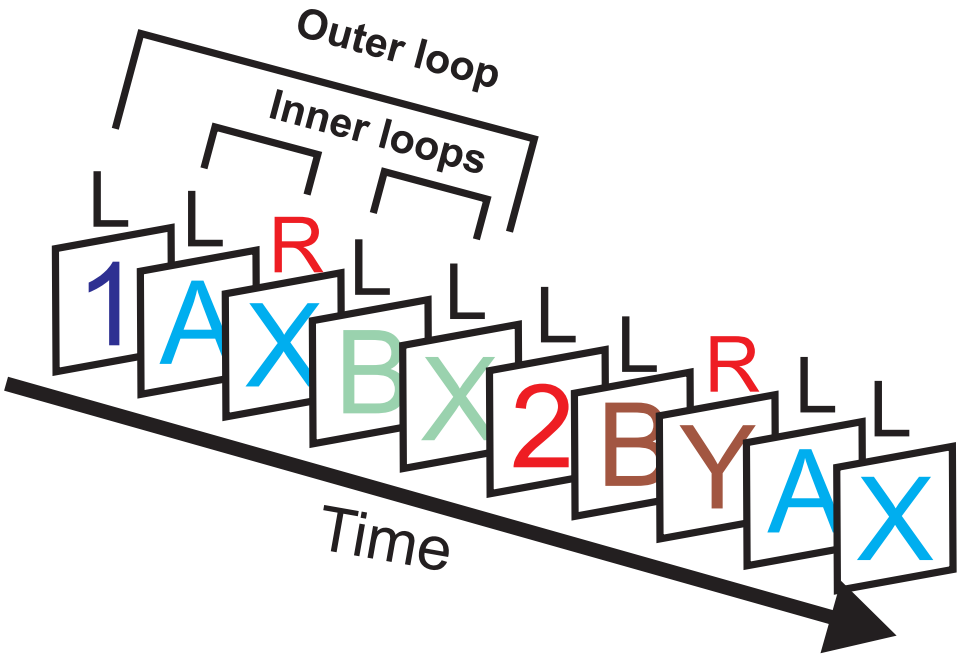


Figure 3.5: The 12-AX task: A linear sequential stream of stimuli is presented with one stimulus shown at a time. Agents need to respond to each with either "L" or "R" . The majority of stimuli are non-targets requiring the default response "L". Only the inner loops `A` immediately followed by `X`, or a `B` followed by `Y` may be a targets ("R") . However, only one target sequence is active at any one time, switching based on the identity of the context markers (`1` and `2`) presented at the beginning of each outer loop. – The colours of the stimuli are only for clarifying the figure and not part of the task itself. Adapted from Frank et al. (2001a).

- Simplicity of the task: For easier modelling of the task, it should be as simple as possible while still providing the necessary properties described above and below.

- A comparison to non-shaped learning: To be able to easily identify the computational and behavioural properties and advantages of shaping, the full task must be sufficiently simple for the learning agent to be able to learn it in its entirety.
- Temporally extended tasks: Although not strictly necessary for the concept of task sequential learning, the use of a temporally extended task provides a typical setting of more realistic cognitive tasks. This introduces the concept of temporal credit assignment during learning on top of any potential structural credit assignment. Furthermore, cognitive tasks are frequently in the domain of partially observable decision making tasks (e.g. POMDP), in which the current input stimulus is not sufficient to uniquely identify the correct output reaction. Instead, the cognitive manipulation of internal state (or working memory) is required to solve the problem. Both of these properties provide additional computational options to benefit from shaping, as shaping in tasks requiring working memory or internal state provides the opportunity to transform the more difficult temporal credit assignment to a simpler structural credit assignment.
- Identical stimulus stream in stages of shaping: In order to maximise the effect of catastrophic forgetting, and thus the effectivness of any architectural support to over come it, each subtask should reuse the same stimulus space. This guarantees that overlapping representations will interfere with each other in sequential learning.

One task that provides all of the necessary properties listed above is the `12-AX` task initially introduced by Frank et al. (2001a). It is a hierarchical extension to the AX variant of the Continuous Performance Task (AX-CPT) (Rosvold et al., 1956), which is itself a popular task for probing executive function and working memory. Furthermore, the `12-AX` task has a strong history in computational modelling of cognitive learning (Frank et al., 2001a; Hazy et al., 2007; O'Reilly and Frank, 2005; Zilli and Hasselmo, 2008; Dayan, 2007; Todd et al., 2009) allowing the results to be compared to a number of different models and linking the conclusions with the larger literature of cognitive modelling.

In the `12-AX` task (figure 3.5), subjects see a sequence drawn from an alphabet of the eight symbols `1,2,A,B,C,X,Y,Z`; every symbol has to be followed by a response. The 'target' key ('R') must be pressed for symbols defined as targets by the rules of the task, and the distractor key ('L') for all other symbols. There are two different inner loops, both of which are 1-back tasks: subjects must declare as a target *either* `X` when preceded by `A` (*ie* to the segment `AX`) *or* `Y` when preceded by `B` (`BY`). These pairs appear without warning in a stream of uniformly-selected random distractor pairs. Every symbol not

defining the end of a target pair should be declared as a distractor. The outer loop is signalled by the numbers, with a `1` meaning that the `AX` task should be performed until the context marker; and a `2` meaning that the `BY` task should be performed instead. The numbers `1` and `2` themselves should also be declared to be distractors ('L').

In addition to these basic sets of rules that govern the correct response given any sequence, there are many possible subtle variants of this task, defined by the statistics of occurrence of each of the stimuli in the sequence. Examples of these variations are e.g. how often are context markers shown, how often are target sequences present or potentially any other structure in the distractor patterns. The statistics originally defined in O'Reilly and Frank (2005) are as follows: Here, each random pair consists of one of {`A`,`B`,`C`} followed by one of {`X`,`Y`,`Z`}, and there are $n = 1 \ldots 4$ pairs in each outer loop. The outer loops are equiprobably `1` and `2`. At least 50% of the inner loops consist of potential target sequences `AX` or `BY`. The other 50% are drawn uniformly from all 9 possible inner loop sequences. An epoch is somewhat arbitrarily defined as 25 outer loops. It results in 150 stimulus presentations on average, with a minimum of 75 stimuli per epoch (if by chance all outer loops only have one inner loop) and a maximum of 225.

Chapter 4

A simple model of shaping and its effects on flexible cognition

4.1 Introduction

As presented in the literature review, researchers have proposed and studied a large variety of models and theories of cognitive flexibility, and the idea of shaping has been around in psychology for nearly a hundred years. However, apart from in robotics and some other examples, the topic of shaping in simulations and modelling of behaviour has mostly been ignored. This chapter therefore provides a computational treatment of shaping in the context of a complex cognitive task, the `12-AX` task, as a proxy for other typical cognitive tasks used to elicit flexible, pre-frontal and basal ganglia dependent behaviour.

In the setting found in most treatments of shaping, such as in robotics and language, prior tasks are always a strict subset of those that follow. Therefore, the subcomponents always continue to be trained. In contrast, the case where several of the intermediate tasks have no immediate overlap was explored, and are only integrated later to form the more complex task. Furthermore, these kind of tasks are learnable within a few sessions at most and contrast to interpretations to explain a form of self-shaping on a developmental time scale (Elman, 1993; Reynolds and O'Reilly, 2009).

The modelling hopes to explore some of the benefits and pitfalls that may be encountered when deploying certain learning algorithms that have been shown to work well in the traditional form of training into the setting of task sequential learning. In this setting, it is assumed that each of the earlier tasks of the sequence (or at least a subset thereof) is an easy task in itself. These can then, at later stages, be combined in some way to form the overall more complex task. Given this potential reduction in task difficulty due to the help of the environment, the architectural, mechanistic elaborations previous models have suggested for achieving good performance in com-

plex tasks, were not used. Instead, by using a simpler, more abstract model, the Long Short-Term Memory model (LSTM) (Hochreiter and Schmidhuber, 1997) (described in detail in section 3.2), this thesis can focus on the computational properties of shaping, and where needed, on additional complexity and sophistication necessary to actually benefit from the extra information given by the environment.

One fundamental aspect any learning agent facing a temporally non-stationary environment has to address, is the issue of catastrophic interference (French, 1999; McCloskey and Cohen, 1989). This results from the stability-plasticity dilemma (Grossberg, 1980; Carpenter and Grossberg, 1988) and is the effect of forgetting earlier training examples by overwriting them with current examples. Indeed, catastrophic interference plays a major role in these simulations, too. Without counteracting it, with the help of a specific mechanism in the network, shaping would worsen the model's performance. This mechanism, presented in section 4.4, is a type of resource allocation. Therefore, the results are mostly in line with the previous research on catastrophic interference. But unlike McCloskey and Cohen (1989), who also attempted a resource allocation as a proposal to overcome the issue in the example of learning arithmetic facts, the current mechanism mostly successfully allows for sequential learning.

Cognitive flexibility often manifests itself in the remarkably short time an agent might only need to adapt to new situations and tasks. For this reason, the main metric used throughout the simulations to determine how successful agents were in benefiting from a shaping protocol, was the average number of stimulus presentations needed to achieve a certain performance goal. In these simulations, the popular `12-AX` task (explained in detail in section 3.5) was used. Comparisons of how quickly networks learnt in the usual way, i.e. immediately trained on the full `12-AX` task, compared against those trained with a shaping protocol, were created. However, speed of learning is by no means the only metric of flexibility. Thus, a battery of additional simulations probing further the differential behaviour between the two training methods was used. Of particular interest was the question as to whether shaping leads to altered internal representations of the rules, and hence affects the way networks operate even after they have all been trained to the same performance criterion.

4.2 Defining a shaping protocol

Key to shaping is identifying the essential sub-components in a task. These can be separated to produce an appropriate training sequence. In case of the `12-AX` task used here, the two main hierarchical components of the task are: i) learning to memorise `1` and `2` for long periods as context markers; and ii) learning to memorise the `A` or `B` for one step to perform the `AX` or `BY` blocks correctly.

Although the exact details of the shaping protocol are somewhat arbitrary, they do

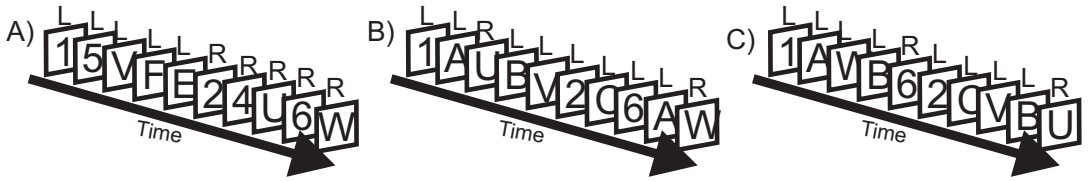


Figure 4.1: Typical sequences used during the shaping procedure. a) The first section of shaping trains the context markers `1` and `2`. Responses to all stimuli following a `1` should be R and to those following a `2` should be L. The maximum length of distractor sequences gradually increases in 4 stages. b-c) The second section trains the unconditional gating of `A` and `B` respectively in a one-back task. The subject has to respond unconditionally with an R to any input immediately following an `A` in b) and `B` in in c).

adhere to the above principles. As shown in figure 4.1, a 7 stage shaping procedure divided into 3 main sections was considered, each of which reflects the structure of one of the sub-components: (i) learning to store the context markers `1,2` (stages 1-4); (ii) learning the one back characteristics of the CPT-AX (stages 5 and 6); and, finally, (iii) the full `12-AX` task (stage 7).

In the first section of shaping, the networks were exposed to a task whose response is only defined by the last seen number. That is, all stimuli following a `1` required an 'L' response, whereas those following a `2` require an 'R' response (see figure 4.1a). The alphabet of possible intermediate stimuli is chosen to be a distinct set of nine further inputs that are not part of the standard `12-AX` task, in order not to confound learning the storage of `1` and `2` with other aspects of the task. Note, however, that in other experiments (not shown) which used the same alphabet here as for the full `12-AX` task, ultimate performance was essentially the same, so this is not a crucial parameter of the procedure.

Gating in the LSTM model is graded rather than binary, partly to ensure differentiability of the network. Making this gating be sufficiently strong required the use of at least some long sequences, in this case including up to 60 intermediate stimuli between successive context markers. Under the shaping procedure (and indeed rather deeply embedded in the strong asymmetry in the task as a whole between the frequency of distractor and target responses), this produces very long stretches of identical responses, which themselves harm learning. Thus, to achieve acceptable performance, training was segmented by presenting this section of the task in four parts (defining the first four stages of shaping), with loop lengths increasing from 12 up to the maximum of 60. Within each individual stage, the distribution of lengths followed a (renormalised) truncated exponential distribution, with the longest possible sequence being more than five times less likely than the shortest.

The second section of shaping consisted of two stages, each based on remembering one of the first elements of one of the two target sequences `AX` and `BY`. Unlike the full `12-AX` task, during these two stages any symbol following an `A` or a `B` respectively required

a target response. Typical sequences are shown in figure 4.1b and 4.1c. The final section (the seventh stage of shaping) involved learning the full `12-AX` task. Overall, the shaping protocol drew upon both facets of shaping. Task sequential learning was used in stages 5, 6 and 7, while stages 2,3 and 4 used a progressive steps approach.

Each stage was trained until the network achieved the performance criterion on that stage, or for a maximum of 500 epochs. This limit was implemented both for computational reasons as well as to remove the odd outlier network that failed to learn in a reasonable time. This limit corresponds to about three times the mean training length, equivalent to several standard deviations larger than the mean of the unshaped base case.

4.3 Basic performance on the 12-AX task

As a baseline for performance, a standard (naive) LSTM network was trained directly on the full `12-AX` task. On average, it acquired the task in 186 epochs (standard error: 8.5 epochs). This was rather faster than the roughly 350 epochs that O'Reilly and Frank (2005) suggested for an unembellished LSTM network. Since the number of memory blocks (4), cells per block (2) and learning rate (0.1) were similar to those they used, this presumably resulted from the altered learning criterion (see 4.3.1) and network parameters.

The choice of 4 blocks with 2 cells was arbitrary, although it did to some degree appear to provide a performance sweet spot between the minimum necessary of 2 blocks and some of the larger networks simulated. As shown here and in chapter 6, performance, however, did not overly depend on these parameters.

By comparison, after complete shaping, networks learnt the full `12-AX` task rapidly, on average in 39 epochs (standard error: 6.6), thus showed the expected large decrease in training time. As the mean can be corrupted by a few outliers, the typical learning times (median) may be more informative. The median for the shaped networks was 14 epochs; for the unshaped network the median was 162 epochs, closer to the mean. Thus, median training times showed even greater advantage with about a 10 fold decrease through shaping.

Of course, the full time for training should also include the time devoted to the shaping itself. Calculating the equivalent number of epochs for the shaping stages based on the number of stimulus presentations, shaping took an equivalent of 74 epochs on average (median is 61 epochs). Thus, as can be seen in figure 4.2, there remained a substantial overall benefit (significant at $p < 0.001$) for shaping. Figure 4.3 shows detailed (averaged) learning curves of each of the individual stages of shaping.

To prevent any occasional bad runs from unduly corrupting average training times, outliers in the form of runs failing to learn any of the stages within 500 epochs were

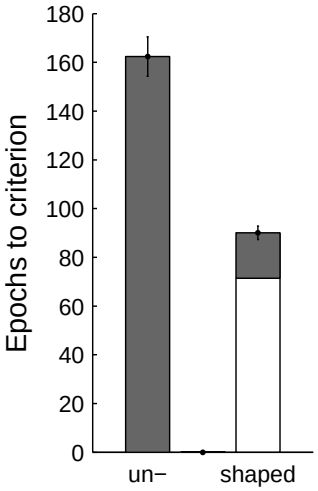


Figure 4.2: Comparison of learning times between shaped and unshaped networks for the standard `12-AX` task. The bars show the numbers of epochs needed to learn the task up to a set performance criterion. Grey represents the epochs trained on the `12-AX` task. For the unshaped network this was the sole training received. White quantifies the cumulative training time of the shaping procedure in terms of `12-AX` epoch equivalents. Error bars represent the standard error of the mean.

excluded from further analysis. This happened for both the shaped and unshaped network in about 5% of the test runs performed to gather statistics.

4.3.1 The influence of the stopping criterion

O'Reilly *et al* (Frank et al., 2001a; O'Reilly and Frank, 2005) defined a model as having successfully learnt the task if it makes no errors in two consecutive epochs. In the simulations, this was found not to be sufficient, as a substantial number of errors can be made even after reaching this criterion. Instead, a softer, but more prolonged criterion, was used, requiring networks to make no more than 5 errors in 30 consecutive epochs, reducing the error rate to 0.5 errors in a thousand responses. These 30 epochs were excluded from the reported training times. Experiments were repeated 100 times with different random weight initialisation and stimulus sequences each time.

4.3.2 Robustness to irrelevant additional structure

The `12-AX` task is the first task seen by the network, which is another potential confound, since, in reality, this task is likely to be only the latest in a very large set of tasks that the subjects will face. The possibility of generalising from these previous tasks to produce even better performance is an important, but hard, question that is discussed later, but cannot yet be simulated. However, it is straightforward to test the robustness of the learning procedure to extra irrelevant behavioural units associated with other, independent, tasks.

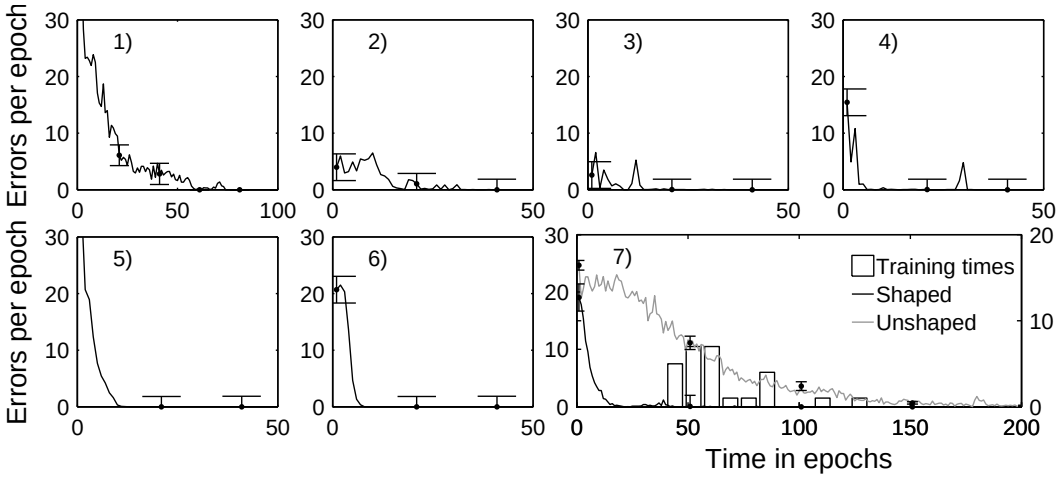


Figure 4.3: Graphs show average learning curves during the individual stages of shaping. The top row shows section one training of the context markers, each presenting training times of increasing length of 12, 25, 40 and 60 respectively. The bottom left two plots show training on the unconditional one-back `A / B` tasks. The last plot shows training times on the full `12-AX` task. The dashed line shows average learning for the unshaped network. The histogram (scale on the right) visualises the distribution of times required for the 6 stages of shaping. Plots follow each other sequentially. Error bars represent the standard error of the mean, and are only shown for selected points for clarity.

Two possible confounding structures were considered, both involving extra, irrelevant, LSTM memory units. Shaping remained identical, with all extra modules fully disabled during the first six stages of `12-AX` task shaping, but enabled and plastic during the final, complete, `12-AX` stage. One set of memory units came from solving a similar task defined on the same alphabet of symbols, but with different context and block markers. The other came from using LSTM modules with random weights drawn independently from exponential distributions, matched to the marginal distribution over weights as occurs during normal `12-AX` training. Distributions were matched separately for the different classes of weights.

In neither case did learning performance using shaping of the `12-AX` task differ much from that of the simple shaped network (means 25, 19 and medians 12 and 10 respectively for similar and random structures).

4.4 The necessity of resource allocation

It has long been recognised that connectionist style models can have severe issues with sequential learning, such as shaping, leading to interference and forgetting. The present simulations were no exception and as shown in 4.4.2 the LSTM also suffered the same problems without additional architectural support. In line with several of the earlier proposals to overcome interference (see section 2.3), the current solution involves a form of segregation to reduce the overlap in the weights. It is called *resource allocation*, as

it implies slow growth in the topology of the network, making new resources or weights available for learning. Reflecting the differences of the LSTM and accounting for the specialities of the 12-AX task, this solution, however, is more deliberate than many of the other attempts.

Especially the temporal pattern differentiates this proposals from earlier ones, where segregation occurs at the resolution of tasks rather than per stimulus. The deliberate nature of the allocation, with a separate mechanism controlling allocation, though, fits nicely with the overall framework of the gated working memory models. It can be seen to extend gating hierarchically to a longer time scale, this time selectively gating plasticity (Sloman and Rumelhart, 1992), rather than selectively gating activity in the network. The details of this allocation are described below.

4.4.1 Manual allocation

Shaping involves separate phases of training sub-tasks. Throughout the experiments, new network resources, i.e., new memory blocks, were allocated by hand when encountering new sub-tasks, ensuring the necessary separation of learnt behaviour across shaping stages. Memory blocks in the LSTM network are mostly independent, and so can readily be treated as separate units of resource. In section 4.9, a method for automatically allocating new blocks is discussed.

For the rest of the results, a single fresh memory block was allocated (having random initial weights) for each new sub-task during shaping. At the same time, previously used memory blocks were temporarily disabled. This resulted in a strict separation of "behavioural units." Each time the allocation changed, weights from the input layer and the recurrent memory layer to the output layer were reset to random values to encourage learning into the new module. In the final stage of learning the complete `12-AX` task, all four memory blocks (three of which were present during shaping and a fourth empty block) were fully enabled, and all weights were plastic. This allowed a fair comparison between the learning of the shaped network and learning of a topologically-identical, but randomly-initialised, unshaped network.

4.4.2 Performance without allocation

As described in subsection 4.4.1, the resources of the network, *i.e.* the LSTM modules, were allocated by hand during shaping, to ensure separation of the "behavioural units". Allocation might be expected to play a critical role in solving the stability-plasticity dilemma (Grossberg, 1980) inherent in shaping, and so for performance without it to be severely impaired. Indeed, the idea of resource allocation, (or more broadly, the concept of increasing the capacity of a learning agent in the event of sudden sequential change in the task to prevent interference of learning) has been proposed previously as

a solution to similar problems. One method was Redish et al. (2007)'s reinforcement learning model of extinction and renewal learning, which is discussed below.

One way to test the importance of resource allocation for shaping is to perform the same learning procedure (the shaping stages followed by the full `12-AX` task), but with all the memory modules being fully active throughout learning. Doing this on average required 235 epochs (median 227 epochs) for the final stage (the `12-AX` task) alone. This was longer, not only than the network which did involve resource allocation, but also than the baseline case of the unshaped network. This problem was only partially solved by increasing the capacity of the network, which should reduce the pressure to find one specific solution. Testing a version with twice as many memory blocks showed a slight reduction in the number of epochs for the final `12-AX` stage (mean 210, median 200). But doubling the number of memory units once more to give four times the initial capacity, led to worse results again.

4.5 Scaling behaviour

The ability to cope with increasingly complex tasks and to identify and learn patterns even over extended time periods is a key aspect of cognitive flexibility. However, long temporal credit assignment paths render rapid learning infeasible for many traditional neural network learning algorithms. By contrast, it is predicted that a network containing the sort of additional task information associated with learning through shaping, should experience fewer problems.

In this case, the most important form of temporal complexity is the number of inner loops contained within each outer loop. The prediction about the benefit of shaping can therefore be tested by training both shaped and unshaped networks on `12-AX` tasks with outer loops varying in length from 1 to 40 sequence pairs, i.e. 2-80 interleaving stimuli. Since the more complex tasks took longer to converge, the cut-off criterion was also raised from 500 to 1000 epochs. All other parameters, including the shaping procedure, were identical to those described above.

Figure 4.4 confirms that, whereas the training time of the unshaped network rose very steeply with longer outer loops (up to 715 epochs on average for 40-length outer loops), the shaped network actually used *fewer* epochs. This apparently paradoxical result came from the fact that the number of pattern presentations per epoch increased (from 75 to 1050 on average). This showed the unshaped network in a particularly unfavourable light.

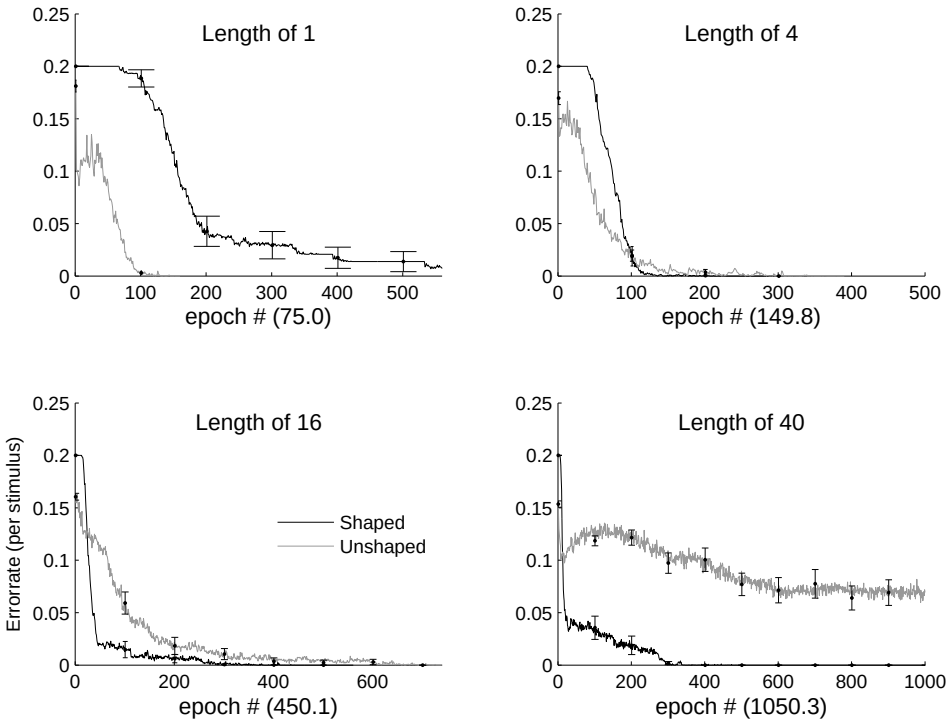


Figure 4.4: Shaping as a function of complexity: Each plot shows learning curves for the `12-AX` task with loop lengths of 1, 4, 16 and 40. Training times of the shaped network are offset by the number of steps (in average epochs) needed to learn the complete shaping procedure. Average numbers of presentations per epoch are indicated in brackets below each plot. The performance of the unshaped network is shown by the dashed line, the shaped network by the solid line. Error bars represent the standard error of the mean.

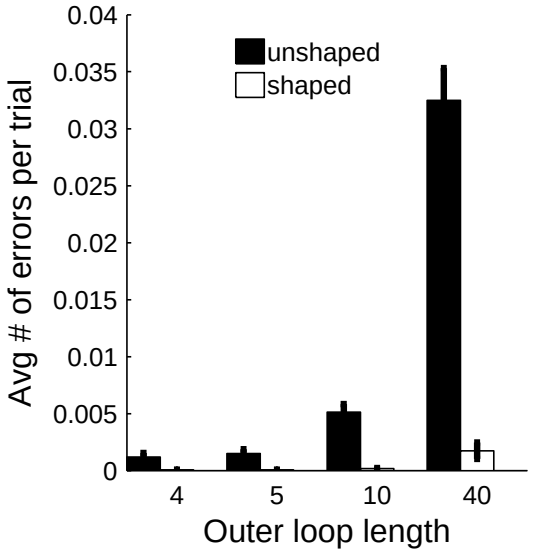


Figure 4.5: Rule abstraction: Shaped and unshaped networks were trained on the base `12-AX` task with a maximum loop length of 4. Without further training, performance was measured on the extended `12-AX` tasks. The loop length of 4 was included as a control and represents the successful training criterion. This graph shows the average number of errors per presentation on each of the different loop length tasks.

4.6 Computational generalisation

Another critical issue is the ability of the procedure to create appropriate computational mechanisms that can generalise along appropriate dimensions by effectively abstracting the statistics of the task away, focusing only on the underlying rules.

In this case, the key dimension is the length of the outer loop. So the ability of the networks, having learnt from outer loops of one length (up to 4 inner loops), to generalise to longer outer loops was tested without further learning. The rules remained the same for each of these tasks. Thus, if they were represented abstractly, in a way that generalises proficiently, then the only extra errors should come from the requirement to retain working memory for more extended periods. Therefore, this parameter manipulation acts as a proxy of one type of rule abstraction. Note that this is a quite different question from that in section 4.5, which concerned learning rather than generalisation.

Figure 4.5 shows that the error rate of the unshaped network increased much more sharply than that of the shaped network, particularly for long outer loops. Figure 4.6 breaks this down by the number of inner loops since the last context marker (*ie* '`1`' or '`2`'). The devastating sensitivity of the unshaped network to this factor is starkly evident. Shaping created a more abstract computational mechanism that generalised appropriately beyond the original training.

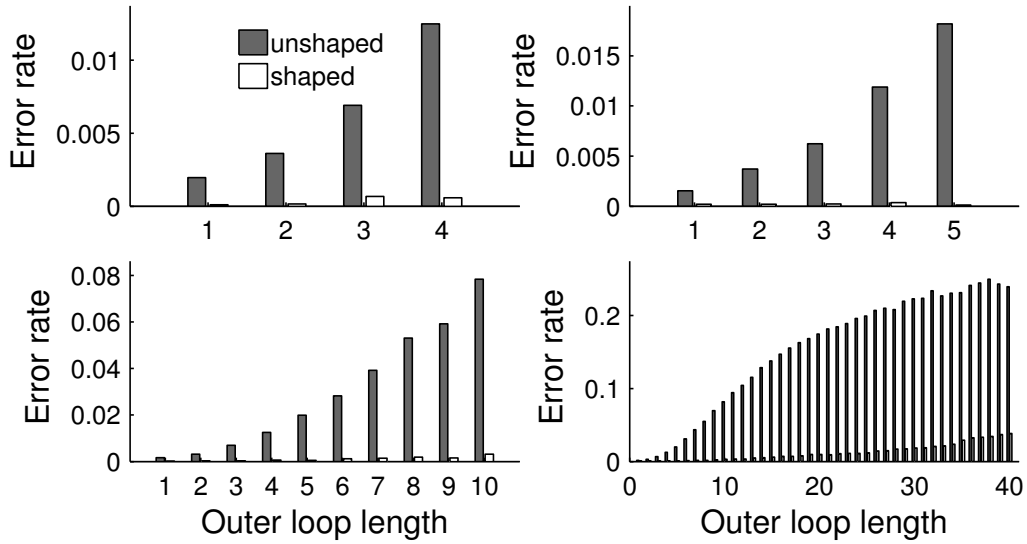


Figure 4.6: Rule abstraction: The graphs show the probabilities of making a mistake on a stimulus n duplets into the outer loop (normalised to the number of presentations of each kind) for shaped and unshaped networks trained on outer loops of lengths 4, 5, 10 and 40.

4.7 Symbol generalisation

A more direct measure of the generalisation capabilities of the network is its ability to handle previously unseen test sequences. To test the shaped and unshaped networks in this way, two of the inner loop sequences during training on the base `12-AX` task were withheld. After successful training to criterion on this reduced task, the error rate on the full task was measured. In order to determine a reliable average error rate, learning in the network was disabled for all cases during the test phase.

As there is only a very limited number of sequences (nine) altogether, the choice of which sequence is left out can significantly influence the results. Obviously the two target sequences `AX` and `BY` cannot be left out, as these are unique and therefore not possible targets of generalisation. Furthermore, each stimulus is only present in 3 sequences. Hence, if two of these containing the same stimulus are withheld, then again, no generalisation is possible. Therefore, it was chosen to withhold `AZ` and `CX` . Even so, only two exemplars were left according to which rules can be inferred, making this test quite hard. With `AZ` and `CX` withheld, the remaining non-target sequences were `AY CY CZ BX BZ`. Therefore, when looking at the `AZ` sequence, the `A` appeared once in a target and once in a non-target sequence, whereas the `Z` never appeared in a target sequence. The `CX` sequence in contrast had the opposite weighting, with the first element of the sequence, the `C` never being part of a target and the `X` being part in half of the sequences. This choice of withheld patterns therefore allowed us to identify if there were differences as to which element was more important for generalisation to a network.

The results of this experiment are presented in figure 4.7. First, when trained on the full 12-AX task, i.e., with no pattern being withheld during training, neither shaped nor unshaped networks had large variations across individual inner loop pairs, showing they had indeed learnt the task correctly. This was to be expected. By contrast, those networks trained on the restricted training set showed more substantial variations. Although both training methods ended up performing worse on the withheld data, either showed generalisation and on average only made on the order of one in 10 mistakes for the withheld sequences. Nevertheless again the shaping seemed to benefit and the overall error rate was down compared to the unshaped training. More interestingly however, the patterns of errors on the withheld patterns differed significantly between the shaped and unshaped networks. Whereas the unshaped network frequently wrongly generalised the CX pattern as a target, shaping reduced this type of error. In contrast however, the shaped network generalised worse on the AZ pattern, occasionally classifying it wrongly as a target. Here the unshaped networks had no problems.

By symmetry, withholding sequences BZ and CY should led to the same result. Indeed, shaped networks made more errors on BZ (shaped: 2.5%; unshaped: 0.0%) whereas the unshaped networks made more errors on CY (shaped: 3.4%; unshaped: 11.4%).

These results suggest that the training method influenced the nature of generalisation, with the shaped networks over-emphasising the first element of the sequence and the unshaped networks generalising along the second element of the sequence. The extensive pre-training on A and B during shaping lead to a prominent representation of these stimuli in the network, and so determined the nature of the generalisation.

To verify this interpretation, a second set of experiments in which a different class of sequences was withheld (AY and BX) was run. By contrast with the previous sequences, both the first and second stimulus of the sequences were each part of one of the target sequences. Thus, from the analysis above, one would expect both types of networks to have problems with generalising these sequences. However, differences could still be seen coming from the nature of the context dependency of the 12-AX task. As shown in figure 4.7b), when splitting the errors by context, the varying types of errors can easily be seen. The shaped network again generalised along the A and more often incorrectly attributed the AY a target in the 1 context and the BY in the 2 context. Conversely, the unshaped network more often incorrectly responded to the AY in the 2 context and the BX in the 1 context. Overall, for both networks, the error rate was much higher with this set of withheld sequences, as each part of the sequence during training was only part of one target and one non-target sequence, making generalisation more ambiguous.

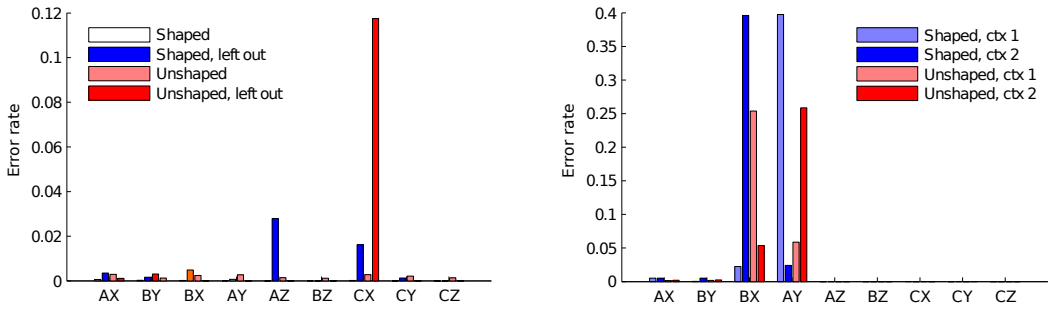


Figure 4.7: Generalisation: These graphs show symbol generalisation in the form of steady state error rates after training on a reduced set of candidate sequences. a) candidate sequences `AZ` and `CX` were never presented during training. For comparison, the results of the fully trained `12-AX` are included. b) candidate sequences `AY` and `BX` are not shown during training. Errors are separated according to the context in which they occur.

4.8 Reversal learning

An important experimental test of flexibility concerns the effect on the speed of learning of successive alternating reversals in the contingencies in the experiment (Butter, 1969; Iversen and Mishkin, 1970; Jones and Mishkin, 1972). A reversed task (`AB-X1`) was therefore defined, in which the rules were the same, but the context markers and the inner loop target sequences were upside down. The particular impact of reversals was compared with that of alternating learning of a new task called `45-DU` (shifting), whose rules were the same, but involved different, non-overlapping, symbols.

Reversal tasks

This set of experiments involved two new tasks, the `AB-X1` task for the reversal experiment and the `45-DU` task for the shifting experiment. All three tasks share the same rules (either one of two target sequences is active, depending on the most recent context marker) and have the same hierarchical nature as the `12-AX` task. However, which stimulus is a context marker and which belongs to one of the target sequence differs between the tasks. In the `AB-X1` task, the same alphabet (`1, 2, 3, A, B, C, X, Y, Z`) is used, with `A` and `B` being the new context markers and `X1` and `Y2` the new target sequences. The `45-DU` task instead employs a non-overlapping alphabet (`4, 5, 6, D, E, F, U, V, W`) with `4` and `5` the context markers and `DU` and `EW` the respective target sequences. By the fact that the `AB-X1` uses the same alphabet as the `12-AX` task, and the tasks are learnt sequentially, the switch between these two tasks requires in the current framework a certain amount of unlearning or at least suppression of the stimulus action mappings. In contrast the non-overlapping alphabet of the shifting task results in less interference due to the strong external indicator of the used stimulus set. Chapter 6 presents further detail on limits for reversal within the same alphabet and

proposes potential extensions to mitigate some of these effects to achieve a high level of flexibility, nevertheless.

In order to accommodate these additional tasks, some changes to the network and the shaping procedure were required. First, the new symbols required the input layer to be extended by a further 9 units. Second, in order to allow the shaped network to be able to build upon prior structure during reversals in the same way as in the base `12-AX` task, the shaping procedure was extended to include equivalent training for the `AB-X1` task. The additional shaping was performed using 4 additional memory modules, so the network had a total of 8. For a fair comparison, the unshaped network was also given 8 modules.

The full shaping procedure of components of both tasks was completed before the first exposure to either the `12-AX` or the `AB-X1` / `45-DU` task, by which time resource allocation was finished and all memory cells were fully enabled. At the time of first exposure to the full tasks, the shaped and unshaped network were again identical apart from the structure in the weights established through shaping. Further, as always, the activations of the network were reset at the end of each epoch, whether or not a reversal occurs.

The actual reversal task consisted of 5 reversals between `12-AX` and `AB-X1`. A reversal occurred either after a network had reached its performance criterion of learning the task, or after a maximum of 500 epochs. The time of reversal was not cued in any way.

Reversal results

The large panels in figure 4.8 show averaged learning curves for the first 6 reversals. Networks with either training method showed substantial difficulties with learning the initial reversal of the task. In fact, the difficulties were such that a large fraction of the training runs failed to achieve the learning criterion within the preset number of training epochs (500). Even for the shaped cases this fraction was at about 50%; however, performance of the unshaped network was particularly devastated by the reversal, with close to all (95%) of the networks failed to converge appropriately. In the graphs showing the error rates averaged over all runs, bar those few failing to learn even on the initial `12-AX`, this failure to learn can be seen in the elevated asymptotes; a somewhat orthogonal feature to the steepness of the initial learning. Nevertheless, over the successive reversals, both networks improved in both measures; but the shaped networks still significantly out-performed the unshaped networks throughout each `AB-X1` task. Shaping did not, however, eliminate the path dependence of the learning of the full task (given the lack of resource allocation at this point); there was typically a very fast phase of initial learning each time the `12-AX` task repeats.

In the shifting experiment (`45-DU` task), compared with the above reversals, switching the contingencies did not lead to such poor performance. The small panels of figure 4.8

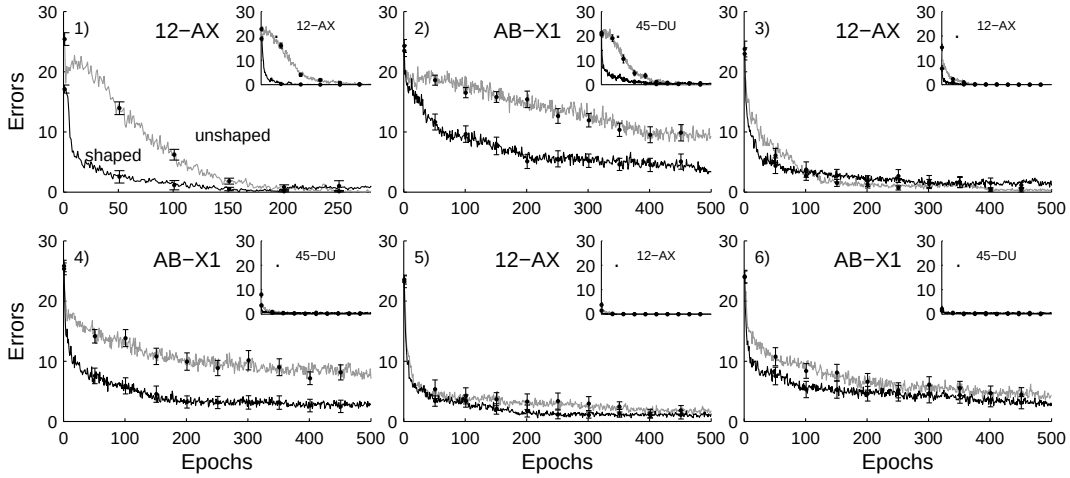


Figure 4.8: Reversal learning (main plots): Each plot in sequence shows the course of acquisition of alternating reversals between `12-AX` and `AB-X1` versions of the task, for shaped and unshaped networks. No changes to the networks and their weights were performed between tasks. The shaping procedure preceding the first presentation of the `12-AX` task is not shown. The inset plots show the equivalent for shifts between `12-AX` and `45-DU` tasks. Error bars show the standard error of the mean.

show that there was a clear learning effect already by the second shift. By the third shift, both shaped and unshaped had mostly learnt the shift and performed well immediately after the switch. However, again, the shaped network did substantially better with the median learning time of 1 epoch for both the last shifts, compared to 14 and 16 for the unshaped network.

4.9 Automatic allocation

In all the simulations so far, a *deus ex machina* was deployed, in which shaping has been supported with the manual allocation of resources at task boundaries. Indeed, section 4.4 showed that shaped networks performed substantially worse than the unshaped networks without allocation. It was done to focus on the potential benefits of shaping in computational modelling, rather than particular implementation details. In fact, the degree to which natural learners would also require, or at least benefit from, additional hints from the environment about task boundaries remains an open question in itself. For human experiments, one could expect revealing these change-points to be particularly beneficial. Indeed, in the experiments conducted in chapter 5, subjects were instructed of the change in task. Similarly, in education, one would seldomly move on to distinct tasks without mentioning the change as this might confuse students. However, in animal experiments involving shaping typically few, if any, hints exist as to when task contingencies switch. Thus, if there was no way of realising these benefits without such an external intervention, then shaping would not be a viable solution. Therefore, one simple mechanism for automatic resource allocation was sug-

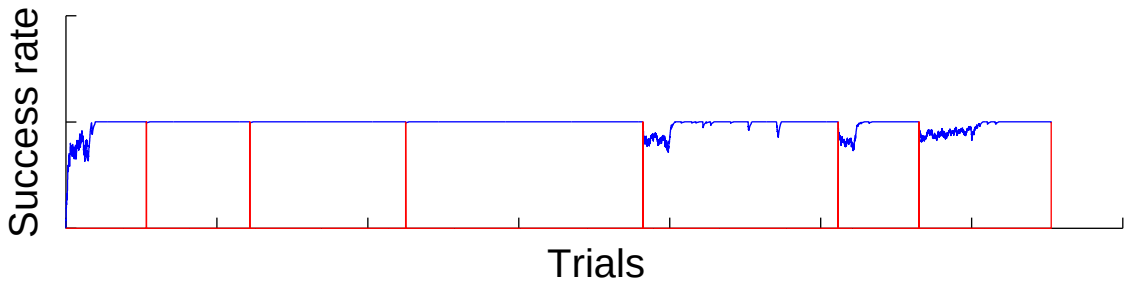


Figure 4.9: Effects of task switching on success rate: This graph shows the filtered success rate of the network as a measure of unexpected uncertainty. The red vertical bars denote the boundaries of the 7 stages of shaping. It shows a clear drop in success rate at the boundaries of the tasks, where an allocation would be needed. The first four tasks are identical other than the length of the inner loop. Thus, no drop in success rate is observed and no allocation should occur.

gested, as a proof of principle. To strengthen the case, not only does it control the resource allocation from within the network, it also uses no additional, external help to detect boundaries at which the task switches occur. The manual allocation results from section 4.3 were considered, and the results on fully automated shaping presented here, to form the extreme cases between which solutions are likely to lie.

4.9.1 Unexpected uncertainty as a trigger for resource allocation

A simple possibility for automatic allocation is to specify a mechanism that detects the onset of each new sub-task. An obvious candidate for this is unexpected uncertainty (Yu and Dayan, 2005; Dayan and Yu, 2006), i.e., the unexpected drop in the performance of the network that happens when each new sub-task is introduced. Indeed, unexpected uncertainty, and its noradrenergic neural representation, have been implicated in the nature and speed of learning in reversal tasks, which involve similar contingencies. Another example of using a form of unexpected uncertainty to detect boundaries is the model of Zacks et al. (2007); Reynolds et al. (2007). In this model of event segmentation in a sequential perception stream, segmentation is based upon the assumption that the predictability of the next percept significantly drops at the boundaries of events, beyond the expected uncertainty within the event, driving a gating signal in working memory. Equally, in the work of Redish et al. (2007), the contingency of extinction is detected by a sudden drop in reward, at which point the state space is duplicated and new resources are allocated.

Along similar lines, the times of heightened unexpected uncertainty was proposed to trigger events in the network. Rather than only affecting network activity through gating, however, a change of the network topology itself (hard resource allocation) was proposed or more realistically through meta-plasticity in the form of differential changes in learning rates (soft resource allocation). This lifts gating, from the level of gating activity to the higher level of selectively gating weight plasticity per module.

For the want of a full model of uncertainty in these simulations, the assumption (which in this case was true) that, having learnt the deterministic `12-AX` task, an agent would make no errors, was made. Thus, the expected uncertainty was close to zero. As such, the current error rate could be treated as a direct measure of unexpected uncertainty. Allocation was modelled as being triggered by a threshold crossing of unexpected uncertainty or error rate increase, a technique both Redish et al. (2007) and Reynolds et al. (2007) employed in their models. As in section 4.4, when allocation occurred, a single new module of the LSTM got activated. However, unlike for manual allocation, the plasticity of the existing modules was strongly reduced (by a factor of 20) rather than being completely abolished, and indeed these modules continued to contribute to network activity in a normal manner. This was essential for a solution to a task to involve a recombination of elements of the solutions to previous tasks.

In order to allow for the occasional error, expected from the stochastic nature of the network, and particularly errors occurring from less deterministic tasks, the success or error rate was calculated as a smoothed (filtered) running average of the binary feedback provided by the environment. The somewhat arbitrary choice of a truncated exponential low-pass filter was made.

$$\text{suc} = \frac{1}{z} * \sum_{i=0}^{299} c(t-i) * e^{-0.01i} \tag{4.1}$$

where $c(t)$ denotes if the trial at time t was correct and $z = \frac{1-e^{-3.00}}{1-e^{-0.01}}$ is a normalisation constant to constrain the score between 0 and 1. New allocation is only possible once the results of previous allocation have surpassed a success rate threshold of 0.99, and there is then a sudden increase in error rate above 10%. As figure 4.9 shows, this filtered success rate was very well suited to detect eventual task boundaries. Indeed, as shown in figure 4.10, this mechanism was quite successful at detecting the correct times of the switch in task. Here, each of the 7 panels represents one of the shaping tasks, with the last the full 12-AX, and each row depicts a rerun of a new randomly initialised network. As can be seen by the blue stars, showing the times at which the network allocated a new block, in the majority of cases, the allocation occurred at the correct time, although occasional spurious allocations occurred, too.

An allocation step consisted of enabling a new memory block with randomly initialised weights. The learning rate of this new block was set to $\alpha = 0.1$ and the learning rate of all previously allocated blocks was reduced to $0.05 * \alpha$ to prevent forgetting of previously learnt tasks.

Although according to the model the number of modules should be able to grow arbitrarily large, for practical reasons it was cut off of new allocation after 12 units, which happened once in the hundred runs.

As there was no outside intervention in the network during the automatic allocation

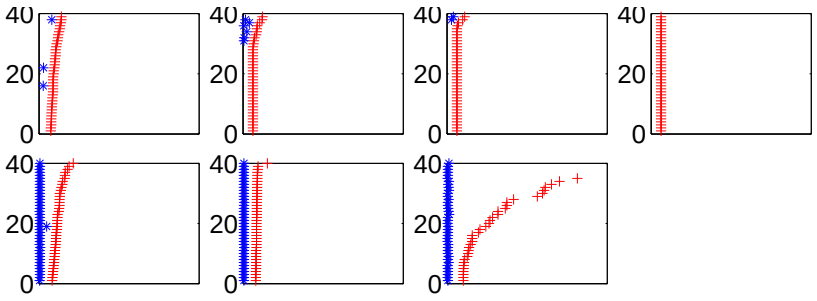


Figure 4.10: Automatic allocation: This graph shows the times at which the automatic allocation mechanism detects a task boundary and allocates a new module to the network. The graph shows 40 reruns (each line representing one) for different random initialisations of shaped training with automatic allocation. Blue stars shows allocation times relative to the actual task boundaries. Red crosses show the point at which each of the networks learnt the respective stage of shaping to criterion. Allocation predominantly occurs right at the switch of task, though some spurious allocations occur, too.

experiments, the network topology at the time of first encounter with the `12-AX` was no longer identical to either the unshaped network or the manually allocated shaped networks. Not only could the number of memory blocks differ due to the continuing allocation of new blocks, but also, once performing the `12-AX` task, not all blocks were equally plastic. Indeed, there was no longer any distinction between epochs involving shaping and those involving the full `12-AX` task.

4.9.2 Results

In order to test the effectiveness of this simple automated allocation mechanism, the basic shaping experiment was repeated. Figure 4.11 shows that this network was a clear improvement on a network without any allocation. Although the mean number of epochs required for learning (173 epochs for the combined training, 87 epochs for `12-AX` task alone) was greater with automated than manual allocation, it was still less than for learning without shaping ($p < 0.05$, single tail t-test). Furthermore, much of this apparent decreased learning speed originated from slightly less robust learning, which made it harder to achieve the stringent performance criterion. This could be seen particularly clearly in the actual learning curve (figure 4.12-7), which differs little from the case of manual allocation in figure 4.3, and showed a large improvement over no shaping. The reduced robustness was apparent in the heavier tail of the learning curve.

Another way of analysing the effects of the automatic allocation is to look at the distribution of points at which new blocks were allocated. On average in each run 3.3 allocations occurred. The majority of runs contained 3 allocations. As 3 allocations are the minimum necessary, the automatic allocation mechanism was well suited at detecting task boundaries. Out of the spurious allocations, the majority of these occurred during the early part of the first stage of shaping, at which point the error rate was

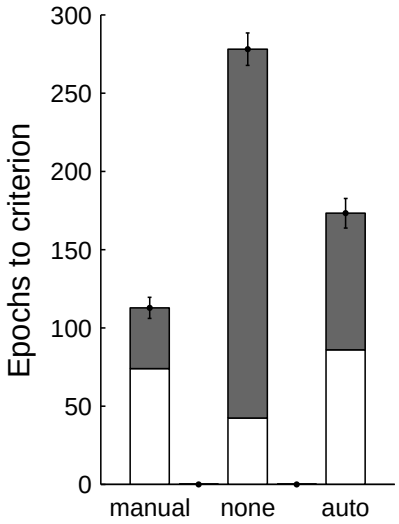


Figure 4.11: Automatic allocation: This graph shows learning times for the shaped `12-AX` task given manual allocation, no allocation, and automatic allocation. As above, the white bars show the epochs devoted to shaping; the grey bars show those devoted to the full task.

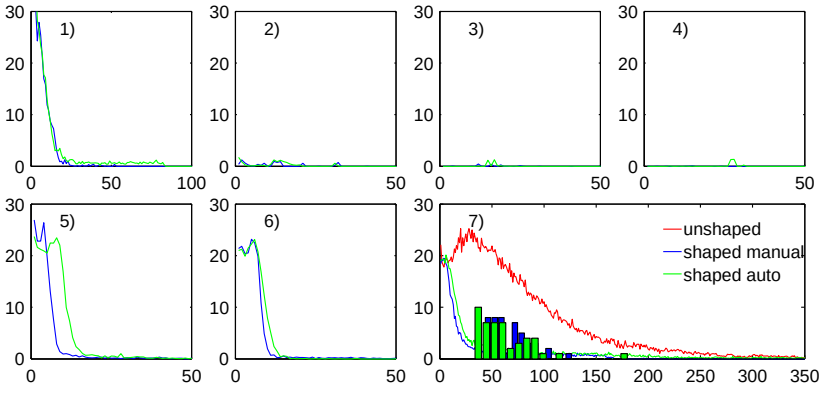


Figure 4.12: Automatic allocation: Averaged learning curves for shaping with uncertainty based allocation. The structure of the graphs is the same as in figure 4.3. Panels 1-6) show the stages of shaping. 7) Learning the full `12-AX`. Error bars represent the standard error of the mean. Histograms show the distribution of combined training times for the shaping stages.

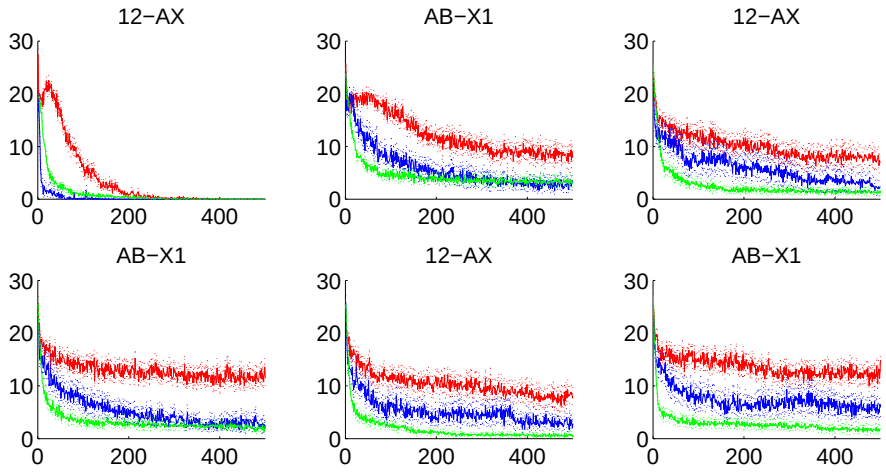


Figure 4.13: Automatic allocation: Reversal learning with the help of allocation. This graph shows the learning curves for an unshaped network, a shaped network with manual allocation and a shaped network with auto-allocation. Standard errors of the mean are shown as dotted lines for each of the 3 learning curves.

still highly variable. Looking closer at the allocation during the final stage of learning the full `12-AX` task, one can see that all runs contained at least one allocation during this stage, and 38 runs detected the task switch within the first epoch of changing to the `12-AX` task.

In a second set of experiments, the reversal tasks of section 4.8 were repeated using the automated allocation procedure. Here, the automatic allocation actually performs *better* than manual allocation. This came from its ability to allocated new blocks throughout the reversals, helping separate out modules associated with `12-AX` and `AB-X1` tasks.

4.10 Discussion and conclusion

Although shaping is widespread as a method of training subjects to perform complex behavioural tasks, its effects in computational models have not been extensively investigated. This chapter studied shaping in a moderately complex hierarchical working memory task, showing that it offers significant benefits for learning in the face of medium-term to long-term temporal demands. Speed is not the only benefit of shaping. It was shown that it also lead to a solution of the task that generalises better over time, and is also more flexible in the face of task changes such as reversals.

There is not yet a clear computational theory that provides constraints on appropriate ways of designing shaping protocols. Indeed, the rapid variation in methods of teaching suggests that there may also be a dearth of clear psychological constraints. It does seem evident that finding the separable hierarchical parts of a problem is key (Watkins, 1989; Singh, 1992; Parr and Russell, 1998; Barto and Mahadevan, 2003). But if there

is more than one way of decomposing a task, or indeed more than two levels in its underlying hierarchy, then further experimentation may be necessary. The possible hypothesis that any way of making a task progressively more complex would be equally valuable was refuted, by showing that training simply with successively longer outer loops without the initial hierarchical decomposition did not show the full benefits, such as generalisation. Further, more subtle changes to such aspects of the procedure as the distribution of inner-loop lengths during shaping could result in slightly different failure modes of shaping. In the future it will be important to try and characterise such differences and test the resulting predictions experimentally.

The simulations showed that shaping alone, without the support of an allocation mechanism, can perform worse than no shaping. This could be because it poses more acute stability-plasticity dilemmas (Grossberg, 1980; Carpenter and Grossberg, 1988). Although section 4.9 presents a simple algorithm for automatic allocation, it is not a complete solution, and substantial further work will be needed to make this procedure robust and general. In particular, to accommodate asymptotic errors in probabilistic tasks, it will be necessary to incorporate expected as well as unexpected uncertainty (Yu and Dayan, 2005). Reynolds et al. (2007) and Redish et al. (2007) showed that using a rapid change in the running average of the error rate can account for some probabilistic effects. However, this fails to accommodate more sophisticated fluctuations in expected uncertainty, and these authors also see a significant drop in performance of their automatic model compared to the manual one.

The obvious alternative is to have the subject explicitly model the uncertainty. In the wider field of machine learning, a popular way of handling the equivalent of allocation involves mixture models, in which underlying modules compete (Jacobs et al., 1991a) to explain inputs based on their own so-called *generative, predictive,* or *forward* models. The MOSAIC framework (Haruno et al., 2001) for mixture model learning for motor control is a good example of this, and could admit a generalisation to shaping. Predictive coding, however, is not only a theoretical construct, but has also been observed in biological system, such as e.g in the visual system (Summerfield et al., 2006; Summerfield and Koechlin, 2008).

Furthermore, even though automated, the resource mechanism remains an additional component to the network and as yet has no direct biological counterpart associated with it. The LSTM network analysis did not aim for direct biological equivalence, as for example the PBWM does. As such no specific areas of the brain that might be responsible for such allocation will be suggested at this point, but the general biological plausibility is considered briefly.

Beyond what the standard models of gated working memory already require and provide, the resource allocation needs the ability to differentially control the plasticity under the influence of the network itself on the time scale of switches in task contingency. This requirement falls well with the definition of meta-plasticity: *"meta-plasticity has*

occurred if prior synaptic or cellular activity (or inactivity) leads to a persistent change in the direction or degree of synaptic plasticity elicited by a given pattern of synaptic activation" (Abraham and Bear, 1996). This phenomenon is known since at least Coan et al. (1989) and has since been observed under various conditions (Abraham, 2008). For example, it is known that meta-plasticity can be influenced by environmental stimuli, as well as induced by learning. It is at least not inconceivable that such a soft resource allocation biologically occurs. However, the use of metaplacticity for "compartmentalising" the network more directly has not been observed. So far it has been implicated with homeostatic and stabilising effects (Bienenstock et al., 1982) or ways to extend longevity of memories (Fusi et al., 2005).

Alternatively, in hippocampus (thought to be responsible for episodic memory, and thus facing the same issue of interference) it is known to be one of the two areas of the brain where new neurons develop even in adulthood (Zhao et al., 2008). Although not directly comparable with the mechanism provided here, it too, presents a form of resource allocation. It has been linked to *"facilitating the encoding of new memories by providing neurons that are not already fully tuned to an existing memory, and therefore are more amenable to learning new information"* and thus *"could facilitate the encoding of new, separated memories while not significantly disrupting older memories"* (Aimone et al., 2010).

Another, related, possibility is to consider resource allocation itself as a recursive instance of gating, now at the level of strategy learning rather than read-in to working memory. In the models, a resource allocation that was exclusive and hence resulted in local representation of tasks was chosen. This type of allocation lends itself comfortably to the simple and abstract model of LSTM. However, these results will hopefully extend at least qualitatively to the more distributed representations that are likely to be found in natural neural networks. The need for explicit resource allocation is also likely to extend to such networks.

In contrast to the type of resource allocation used here (derived from the ideas of mixture models) a common alternative approach to overcome the catastrophic interference, is the idea of memory consolidation during sleep. This concept, for example, is embedded in the Complimentary Learning System model (CLS) and its extensions (McClelland et al., 1995; O'Reilly and Norman, 2002; O'Reilly and Rudy, 2001; Norman and O'Reilly, 2003).

This chapter focused on shaping in operant conditioning. Three additional forms of shaping are also important, and would be interesting targets for future studies. First, in sculpting animal behaviour, it has been observed that it is beneficial, or even essential, to start from the actions intrinsically emitted as Pavlovian responses to predictions. Breland and Breland (1961) provide some striking and memorable examples of this maxim, and, as in negative auto-maintenance (Sheffield, 1965; Williams and Williams, 1969; Dayan, 2006), the deleterious consequences of ignoring it.

The second form of shaping is self-shaping, in which subjects themselves may explicitly simplify tasks, by omitting certain features (Duncan, 1995), either deliberately, or perhaps just by not understanding them. Elman (1993) shows the potential benefits of this in the case of grammar learning. However, explicit self-shaping requires the automated discovery of the hierarchical structure of tasks, which is highly non-trivial itself. Such hierarchical structure discovery has been an active, though not currently strikingly productive, focus of work in machine learning (McGovern and Barto, 2001; Barto and Mahadevan, 2003; Bakker and Schmidhuber, 2004).

Finally, shaping can involve the formation through training of more abstract representations of input that can be used to speed the subsequent learning of complex behaviours. In the field of machine learning, this is the standard view of the interaction between unsupervised learning, which re-represents inputs according to the underlying structure of their statistical distribution, and supervised learning, which uses these representations to perform tasks well (Hinton et al., 2006; Hinton and Ghahramani, 1997). However, it is also possible to generalise the use of representations learnt directly to solve one task to other tasks. For instance, following on from Premack (1983), Thompson et al. (1997) showed that chimpanzees that had been trained on a sophisticated task of determining the identity of two inputs, putatively building a new, abstract, representational unit, could more readily learn a separate task, in which the identity or otherwise of two initial inputs determined what actions should be subsequently executed.

Chapter 5

Human performance on the 12-AX task

5.1 Introduction

Previous chapters have presented models, mechanisms and ideas potentially underlying the transfer of learning from one cognitive task to the next, and presented some behavioural predictions arising out of these models. So far, these ideas have been based on theoretical considerations alone. Ultimately, however, the aim of this research is to identify mechanisms of cognitive learning as they actually occur in humans. To this end, it will be essential to validate the findings from the present study and others (e.g Frank et al., 2001a; Hazy et al., 2007; O'Reilly and Frank, 2005; Zilli and Hasselmo, 2008; Dayan, 2007; Todd et al., 2009) in human learners. In particular, it will be necessary to confirm that the benefits of shaping observed in the network are recapitulated qualitatively and quantitatively in human subjects. Moreover, it will be interesting to determine whether manipulations in the task parameters result in the predicted alterations in learning behaviour. Finally, fMRI studies of subjects during shaping will provide important insights into the neural networks underlying these behaviours in vivo, which in turn will be instructive in the further refinement of various models. For all of these reasons, a detailed analysis of human performance during shaping on the 12-AX task will be critical to our research in the future.

Surprisingly, however, virtually no information exists to date on human performance on this task (see below for a description of some details known thus far). Despite its popularity in modelling, the full 12-AX task has so far not been probed in any behavioural learning experiments in either animals or humans. It has been unknown if humans are capable of learning the full 12-AX task in a trial and error paradigm, and particularly to what degree and with what speed. It is also unknown how models of learning reflect actual behaviour on this task. This chapter begins to explore these

important questions. It aims to determine if 12-AX is a viable task to use for testing predictions made by various researchers' models, including the ones presented in this thesis. Furthermore, it attempts to begin to explore the many free parameters in this task, such as the stimulus presentation statistics or the pictures used. By identifying how some of these affect the difficulty of the overall task and subjects' ability to learn, it aims to present some guides for how these may be chosen in future experiments. In addition, given the focus of this thesis on shaping and the transfer of knowledge to new tasks, it tests the hypothesis that humans are indeed able to take advantage of the task being broken up according to its hierarchical structure, with each component presented separately prior to training on the full task. Based on the results obtained from this experiment, preliminary evidence is provided that the model indeed recapitulates selected elements of shaping and cognitive flexibility in humans, although significantly further extensions will be required to the models before the full complexity of human rule learning is captured.

In the only existing study on the full 12-AX task, Reynolds (2007) presented preliminary data on humans executing the task after instruction. He showed that the execution itself poses no great difficulty. Subjects responded more rapidly to the context-dependent expected sequence than incorrect sequences, a finding perhaps akin to those commonly seen in serial reaction time tasks (e.g. Nissen and Bullemer, 1987; Keele et al., 2003; Robertson, 2007). Further, fMRI results confirmed that two of the brain areas involved in the 12-AX were the dlPFC and aPFC, in a way compatible with the various models. The study, however, did not probe learning.

The AX-CPT as a precursor to the 12-AX

Despite the lack of data on the full task, many individual components of the 12-AX have received more attention. For example, working memory, an important aspect of the 12-AX, has been studied extensively in paradigms like the n-back task (e.g. Braver et al., 1997, 2001; Cohen et al., 1997; Gevins and Cutillo, 1993; Rowe et al., 2000), reviewed in more detail in 2.5.1. Study of the AX Continuous Performance Task (AX-CPT), the non-conditional predecessor of the 12-AX, reaches back at least to Rosvold et al. (1956). Here, a forced timing paradigm was used to test attentional affects of brain damage, and whether the AX-CPT could be used to distinguish between brain damaged subjects (not controlling for its type) and healthy subjects. Later, the AX-CPT task was used in the behavioural evaluation of schizophrenia (Servan-Schreiber et al., 1996; Nuechterlein and Dawson, 1984). Schizophrenic patients performed selectively worse in the AX-CPT than in a test requiring a simple response to a target in a continuous stream of stimuli. This effect was attributed to patients' reduced ability to construct and maintain an internal representation of context and was subsequently modelled in a connectionist framework by Cohen et al. (1999); Cohen and Servan-Schreiber (1992)

Recently, Eshel et al. (2009) used TMS in the AX-CPT task to probe the ideas of gated working memory directly. They showed that applying TMS pulses to PFC shortly after the presentation of the stimulus needing to be memorised could disrupt performance. As TMS pulses applied later in time, but still before recall, had less effect, the data suggest TMS had an effect on gating rather than working memory storage per-se.

More recently, the 12-AX task itself has been proposed as a candidate diagnostic instrument in schizophrenia, although later was rejected as *"the breakout group felt that the 1-2 AX-CPT was an interesting and promising task but that it needed more research at both the basic and clinical level"* (Barch et al., 2009a,b) and was subsequently listed as *"No published data"*. It could be that the conditional, or context dependence of switching between the AX and BY elements of the task, reflects even better some of the described traits of schizophrenia such as "incapable of holding train of thoughts in the proper channels", "disconnecting of associative threads" or "keeping their attention fixed for any length of time" (Bleuler, 1911, 1924; Kraepelin, 1913), which was already motivation for using the AX-CPT in the study of schizophrenia.

Context based switching

The central difference between the AX-CPT and 12-AX is context based switching and the hierarchical gating with varying time scales. These aspects alone enjoy a long and varied experimental history, for example in the form of various task switching experiments (Logan and Bundesen, 2003; Monsell, 2003; Hyafil et al., 2009). These have been used to probe cognitive flexibility and assess it in various brain injuries and psychiatric diseases (Meiran et al., 2000). As far back as at least Jersild (1927), it was observed that reaction times are higher after a task switch occurs than after a repetition of the same task. This has been hypothesised to be because of the need for cognitive control. How the current task is cued varies between studies, but two of the most relevant ones are pre-specified task sequences (Allport et al., 1994; Mayr and Keele, 2000) and intermittent instruction paradigms (Gopher et al., 2000). In the first paradigm a whole sequence of tasks is cued and this must then be kept in memory for execution. In the second paradigm, tasks switch randomly as in the 12-AX task, although, typically, task cues are repeated for each stimulus.

One aspect of task switching is suppressing other rule sets. This aspect has been widely studied in the Stroop task (e.g. Stroop, 1935; MacLeod, 1991), researching the involvement of various brain areas (Banich et al., 2000). However, unlike the 12-AX task, where both contexts are equiprobable, the Stroop task is typically used in an asymmetric setting, for which one response is more common than the other. MacLeod and Dunbar (1988) did study the effect of asymmetry including the symmetric setting, by training the two individual tasks to a different degree a priori, showing that interference can be bi-directional and depends on automaticity rather than time of processing,

although both would be expected to be equal in the case of the 12-AX.

Rule and task learning

Nevertheless, in all these studies, subjects were instructed with the complete set of rules. Thus, they only probed the steady-state execution performance of subjects, i.e. their ability to sustain the necessary attention and load on working memory, rather than their ability to learn and identify the rules from trial and error.

Subjects' ability to discover rules has been probed in other characteristic tasks. For instance, in the Wisconsin Card Sorting task (Grant and Berger, 1948), subjects must use trial and error to track an occasionally switching dimension by which to sort a deck of cards. Performance depends heavily on prefrontal structures (Nagahama et al., 2001; Berman et al., 1995), with damage resulting in reduced ability to switch, i.e. patients frequently persevere in the dimension. So the task is often used to assess cognitive deficits of frontal patients (Demakis, 2003; Milner, 1963). In comparison to the 12-AX task however, learning is limited to the discovery of the task relevant dimension between switches. For healthy subjects this is rather simple, as the rule set of the task is otherwise fully explained. It is more related to the steady state behaviour of the generalised 12-AX as described in chapter 6

The most popular experimental paradigm for studying implicit sequential structure learning is the artificial grammar learning (AGL) (Reber, 1967; Gomez, 1997). This is somewhat similar to a serial reaction time task and extends sequential learning from the motor domain to the judgement or cognitive domain. During training, subjects are presented with strings generated from a simple artificial grammar (realised by a Markovian state machine). In testing, subjects then have to judge what new strings are "grammatical". Typically, subjects are not actively aware of the rules and rely on "gut feeling" to answer. Nevertheless, they perform above chance.

Todd et al. (personal communication) have been conducting similar human learning experiments on variants of the AX-CPT tasks. These have also used fMRI on the underlying areas involved in learning. At the time of writing, the results are unpublished. Unlike the single session experiment presented here, they chose to allow learning to extend over 10 sessions on consecutive days. As a variant of the AX-CPT, however, the rules contained a single loop, only, rather than the conditional hierarchical version of the 12-AX. Furthermore, the details of both, stimulus and response presentations as well as their statistics, differ to the present work and this can have marked effects on the ability of subjects to learn, as both studies show to varying degrees.

5.2 Experimental set-up

First the experiments are described and how the abstract concepts described in earlier chapters have been adapted to be suitable for human testing. The methods used to analyse the data gathered are described.

5.2.1 Participants

52 participants were recruited from the UCL psychology subject pool and randomly assigned to one of the 4 experimental conditions (see 5.2.3). Data of 3 participants were excluded due to the experiment ending early and not having a full set of results. Of the remaining 49 participants, 29 were female and 20 male. All subjects were healthy, had no learning disabilities and were over 18 years of age. A majority of participants were college students, although the database is open to anyone living in London, and participants with a variety of other occupations and backgrounds took part. The native languages of participants were also diversely distributed, but all subjects were versed sufficiently well in written and spoken English to understand the instructions clearly. Nearly all participants were well acquainted with typical behavioural and psycho-physics experiments, but were naive with respect to this task. Participants were paid £7.50 per hour for taking part. To ensure they were motivated to learn the task as quickly as possible, an additional bonus of up to £7.50 was paid, comprising of £5.00 if they achieved the criterion during training, and up to £2.50 depending on their test performance. All participants gave written informed consent before participating in the experiment and all procedures were in accordance to the ethics guidelines and approved by the ethics committee of University College London.

5.2.2 General set-up

The experiments were designed to be in close accordance with the task conditions used in the simulations in chapter 4, with a few obvious difference to accommodate for human participants as well as some variations to test for specific aspects (see 5.2.3). The task was programmed in Matlab^{TM} (The Mathworks Inc.) using the psycho-physics toolbox (Brainard, 1997; Pelli, 1997).

Instructions

At the beginning of the experimental sessions, participants were given written instructions explaining the details of the experiments as accurately as possible, without deception. In addition to the general set-up, instructions also included specific information about some of the properties of the tasks to be learnt. Subjects were told that the task rules are fully deterministic and that feedback is always accurate. Due to the very

asymmetric nature of the task, where only a few responses lead to errors, it was important to ensure subjects were aware that any errors made were due to not having fully solved the task, rather than unavoidable errors due to any probabilistic component. Furthermore, subjects were told that the correct response depended on the current stimuli as well as some of the previous ones. No more precise information about the temporal dependencies amongst stimuli was given, as this represents one of the key components participants needed to learn during the task. Finally, subjects were told that their response only influenced the reward received, whereas the actual task continued independently, i.e. that they had no influence on the task progression. This was done to limit the set of hypotheses participants needed to consider, ensuring that solutions lay within the space achievable by the models described in earlier chapters. The instructions for subjects in all experimental conditions included the warning that there may be switches between rule sets. This allowed for the condition of shaping (see 5.2.3). However, subjects were told that rules only switched at the boundary of epochs and only when clearly indicated by an extra screen to instruct them of this change. In addition, they were told that new rule sets were not independent, and thus old ones should be kept in mind. To ensure subjects had really understood the instructions, a verbal run-through was given after the written instructions, and they were allowed to ask questions. While answering, careful consideration was taken to ensure no additional information about the nature of the task was revealed. The exact instructions are listed in appendix A.2.

Phases of the experiment

During experiments, participants sat at a computer (standard laptop with 13" TFT LCD screen) in a quiet office. They were free to adjust the set-up to ensure a comfortable posture and had no restrictions on how they looked at the screen. Stimuli were black pictorial drawings, or letters and numbers (see 5.2.3) on a white 256 pixel square background, roughly spanning 5 cm on screen. While the experimenter stayed during the initial epochs most of the time and occasionally checked to ensure everything was fine, no other person was present in the room.

Each session was split into three phases, the familiarisation phase, the learning phase and the testing phase. The familiarisation phase was very brief. It ensured subjects had seen all the stimuli, could recognise the feedback and understood the general timing of the experiment. For the first 5 stimuli, the experimenter, rather than the subject, pressed the response to demonstrate the experiment practically. After that, subjects would respond to further stimuli in the sequence. Once subjects said they were confident they understood the task, they would progress to the next stage. Although variable, no subject was exposed to more than 20 stimuli during this period, most considerably less and is thus not likely to effect the learning outcome. The majority of the experiment consisted of the learning phase, in which participants attempted to learn to perform

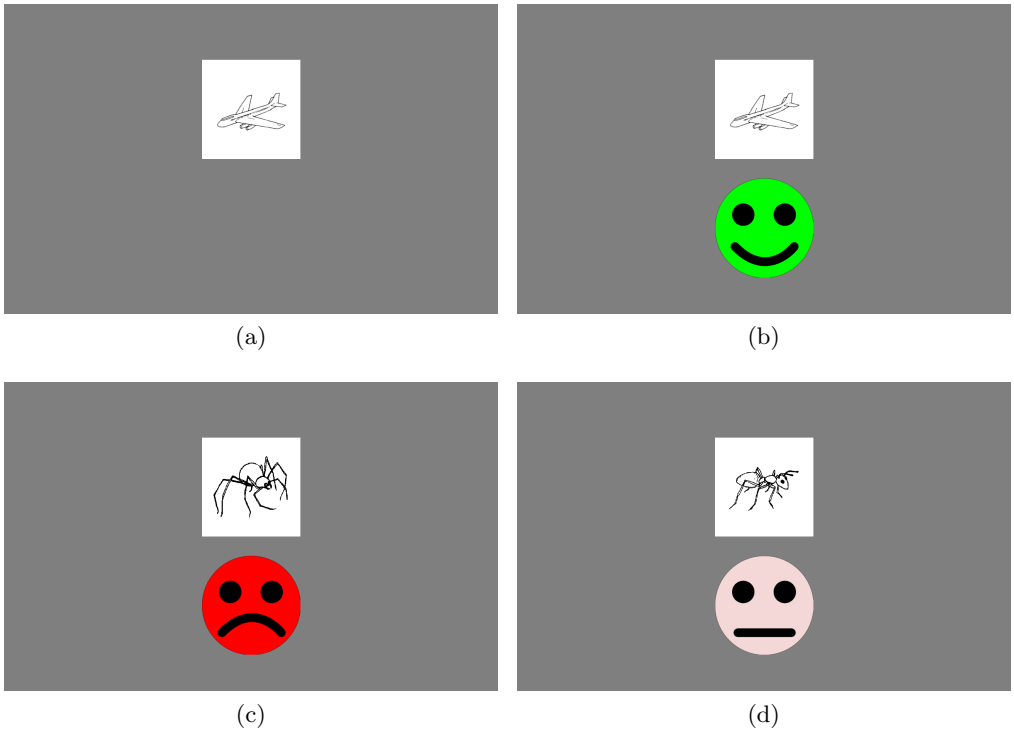


Figure 5.1: Typical screens seen by subjects during the experiment: (a) Subjects were presented with a stimulus to which they have to respond (b) After a correct response, positive feedback was given by presenting a green happy face beneath the stimulus for 500 ms, after which the next trial immediately started. (c) After an incorrect response, negative feedback was given in form of a red sad face. (d) During testing phase, no feedback was given, although the graphical presentation and timing was matched by showing a gray neutral face, independent of the correctness of the response.

the task as well as possible. Once they achieved the learning criterion (see below), they immediately moved into the final stage of testing, which again was short with only 2 epochs ($\approx$ 300 trials). The testing phase was identical to the training phase, with exactly the same task presentation, except that no meaningful trial-by-trial feedback was provided. Subjects could therefore no longer use reward to apply, for example a win-stay, lose-switch strategy. The testing phase thus allowed an accurate assessment as to whether they learnt the rules, rather than using some different strategy. In order to keep the set-up as similar as possible with respect to timing and saliency, "neutral feedback" was given at the appropriate times instead.

During the learning phase, in accordance with the rules of the task (see section 3.5), subjects were presented with one stimulus at a time, as shown in figure 5.1a. They then needed to respond by pressing either the right or left cursor key. Any other key press would automatically be counted as a mistake. Subjects were not time-limited in their response, other than through a long 20 second time-out, which they did not reach under normal conditions (54% of the subjects did not exceed limit once, with a mean of 1.7 timeouts in the entire training session). Immediately after the response, positive or negative feedback was shown beneath the current stimulus in the form of a green happy face or red sad face (see figure 5.1b-c) for a duration of 500ms, after which the next stimulus was presented. After every epoch, i.e. 25 outer loops (see 3.5 for the definition) corresponding on average to 150 stimulus presentations, they could take a short break. Furthermore, at the end of each epoch, subjects were presented with the percentage of the achievable weighted reward (see below) they obtained, to get a hint as to how well they were doing so far. Given that the binary trial-by-trial feedback did not give any hint of the asymmetrical reward rate, this asymmetry could not be used to infer, which stimuli were more important than the others.

Asymmetric reward calculation

The weighted reward measure, that was presented to subjects between epochs and used for analysis in section 5.3, reflected the asymmetric nature of the task. To calculate it, stimuli were split into two categories: Potential targets and non-targets. A potential target consisted of an `X` that was preceded immediately by an `A`, or a `Y` immediately preceded by a `B`, independent of the context (`1` or `2`). All other stimuli, including `X` and `Y` not preceded by the appropriate stimulus, were non-targets. Correct responses to potential target stimuli were rewarded 8.5 times more than correct responses to non-target stimuli. Incorrect responses were charged by the same amount, so that a performance of 50% correct would lead to 0% reward. Subjects were told that the reward rate was not a simple percentage of correct responses, but instead a weighted percentage with the more difficult stimuli being rewarded more. The exact method of calculation and choice of stimulus weighting was not revealed to subjects. Although this reward calculation is fairly arbitrary, empirically it helped spread out the percentages

between those who learnt and those who didn't.

Criterion for successful learning

The criterion used to decide when a subject had learnt the task and could move to the next phase of the experiment, was two consecutive epochs in which at least 90% of the calculated weighted reward was achieved. 90% was sufficiently low to ensure that participants who knew the rules would not have trouble achieving the criterion, despite occasional lapses of concentration. At the same time, two consecutive epochs above 90% was a minimum criterion necessary to ensure that subjects had acquired the full set of rules. The assessment phase and debriefing revealed that, given the probabilistic sampling of stimulus sequences and the relative propensity of targets, single epochs above 90% could very occasionally occur even though participants did not know the full set of rules. With two epochs, instead, all participants achieving this criterion could repeat the performance in the test phase and were able to reproduce the rules in the debriefing session, although they did not always find this easy. Participants were told about the criterion and were encouraged to achieve it as quickly as possible as their ultimate payment depended on it. Independent of performance, the training phase was limited to a maximum of 25 epochs or a total experimental time of two hours. This ensured that the experiments could be conducted in a single session per subject. Participants who did not achieve the performance related criterion in the given time, counted as not having learnt the task.

5.2.3 Experimental conditions

Altogether, there were 4 different experimental conditions in this study, (the `12-AX` task, the Apple-Axe task, the shaped Apple-Axe task and the Spider-Train task). Each is a variant of the basic 12-AX task, i.e. following the same set of underlying rules but manipulating specific aspects to test different hypotheses. Participants were randomly assigned by the computer at the beginning of the experiment to one of these conditions and were not informed as to which version they would face.

12-AX task

The first experimental condition, the `12-AX` version, was the closest overall to the task described in chapter 3.5, using the same parameters as described there. Each epoch consisted of 25 outer loops of context markers. Each outer loop consisted of between 1 and 4 inner loops, drawn from a uniform distribution. Each inner loop had a paired structure, with the first stimulus in the pair drawn from the alphabet {`A`, `B`, `C`} and the second stimulus from the alphabet {`X`,`Y`,`Z`}. The sequences of stimuli presented were drawn probabilistically, but with a distribution such that pairs `AX` and `BY` had

a prevalence of 66%. Context markers 1,2 were equally distributed with 50% chance. The stimuli used to show subjects were literally those of the 12-AX task, i.e. a picture of the numbers 1 , 2 and letters A, B, C, X, Y, Z. They are shown in the appendix at figure A.1. Stimulus assignments were not randomised in this condition and the same stimuli always had the same meaning. This condition was the easiest (in terms of fastest and most reliable to learn), and was used as a baseline as to whether humans can learn the general task at all.

Apple-Axe task

Compared to the 12-AX, the Apple-Axe task had two important differences. The main difference was a change to the stimuli used. Rather than alphanumerical stimuli, the Apple-Axe version used abstract black and white line drawings of simple objects at the level of basic categories, such as an apple or an axe. The stimuli were taken from the standardised data set of pictures developed by Snodgrass and Vanderwart (1980). This data set was explicitly designed for experiments of cognitive processing and working memory and was standardised on 4 relevant variables: name agreement, image agreement, familiarity and visual complexity. Out of the 260 pictures in the original data set, a specific subset of 9 pictures was chosen and mapped onto the alphabet of the 12-AX. These were chosen such that all pictures were equally dissimilar to each other, i.e. came from different top level categories and had no additional structure. The exact set of pictures is shown in appendix A.2. The mapping of pictures to elements of the task, i.e. which stimulus represented the context markers or target sequences, was fully randomised from participant to participant. Equally, the assignment of target / non-target responses to the right and left cursor buttons were randomised. Throughout the results section, the stimuli will be referred to by their equivalent meaning though, i.e. 1 and 2 for the context markers and AX and BY for the target sequences.

A second change was that inner loops did not have the same paired structure as the 12-AX version. That is, rather than drawing a stimulus from {A, B, C} followed by one from {X, Y, Z}, the stimuli between the context markers were each randomly chosen from the full alphabet of {A, B, C, X, Y, Z}, including allowing the same stimuli to be repeated. The probabilities of each of the stimuli were adjusted, so that the overall number of possible targets AX and BY were matched to those in the 12-AX version. In comparison to all other versions of the task, the Apple-Axe task contained no additional cue that could help identify the structure of the task. This version therefore corresponded most closely to the task that the models in previous sections faced in the absence of shaping. This version of the task was the hardest.

Shaped Apple-Axe task

The third condition, the shaped Apple-Axe task, tested whether humans can take advantage of a shaping protocol, such as that used in the earlier chapters. Subjects were shaped using four different rule sets. Each rule set was presented until the learning criterion for that task was reached (although the shaping tasks used the less stringent criterion of only one epoch above 90%, instead of two), or again for a maximum of 25 epochs. Subjects were then moved to the next rule set. Boundaries of rule sets were clearly marked by an additional inter-epoch screen mentioning the change. The statistics of the sequence of stimuli presented were identical in all 4 rule sets, but the mapping from stimuli to correct button presses were different in all cases. These statistics corresponded to those of the Apple-Axe version, using the stimuli shown in figure A.2 without imposing categories and without the pairing structure of the initial `12-AX` condition. In the first stage of shaping, the stimulus-response mapping required a switch in response solely on the context markers. All stimuli in the '1' context, including the 1 stimulus itself, required a left (or right depending on randomisation) press. All stimuli in the '2' context, on the other hand, required the other button to be pressed. This rule set allowed the participant to learn two particular aspects of the task: a) that the rules included switching, rather than just depending on sequences, and b) the identity of the stimuli that caused the switching. The second rule set was based on the inner loop alone. Here, subjects had to respond with the target button whenever they saw AX in the sequence, independent of the context they were in. The last rule set in the shaping sequence was similar to the second, this time using the other target sequence BY. The last two sets were intended to make it easier to identify the target sequences, as they were always rewarded in their respective rule sets, independent of context. The order of shaping rule sets was always the same. The full 12-AX followed as the fourth rule set in the sequence. After completion of the full 12-AX, subjects moved into the testing phase just like in the other experimental conditions. The testing phase only contained the full 12-AX rule set and not any of the shaping sets. Although participants were aware of each individual switch of rule set during the training phase, they did not know how many rule sets there would be, and did not know which rule set would be the one they would encounter in the testing phase until they entered that phase. They were then instructed that it consisted of the task immediately preceding the testing phase.

Spider-Train task

The Spider-Train version was designed to test if humans were able to pick up on additional structure presented within the sequence and use it to learn the task more efficiently. This version used a different subset of 9 images from the full data set of 260 images. Unlike the Apple-Axe images, which were chosen to be equally dissimilar

to each other, the Spider-Train images (shown in appendix A.3) were selected to fall into 3 clearly distinct groups comprising 3 images each. The three categories were animals/insects, music instruments and transportation vehicles. These groups were aligned with the structure of the task, with the context markers 1 and 2 randomly sampled from one group; the first element in the target sequence A, B, C from the next group, and the second element in the target sequence X, Y, Z from the third group. Otherwise, the same parameters were used as in the Apple-Axe version.

5.2.4 Data analysis

Learning curves were calculated as follows. Data for each subject was binned into epochs and the average reward received per epoch was calculated according to the asymmetric reward formula described above (see 5.2.2). Each subject's reward per epoch was normalised by the maximum reward a subject could have achieved in this epoch. This was necessary as the maximum achievable reward per epoch could vary, dependent on the randomly sampled total number of trials per epoch and the resulting number of potential target sequences in the overall sequence. The high variability and asymmetry of task also forced a rather long averaging period of an entire epoch, rather than employ a more fine grained analysis such as the state space smoothed learning curve. (Wirth et al., 2003; Law et al., 2005). The rewards normalised per subject were averaged together for all subjects in a specific version of the task, independently for each epoch.

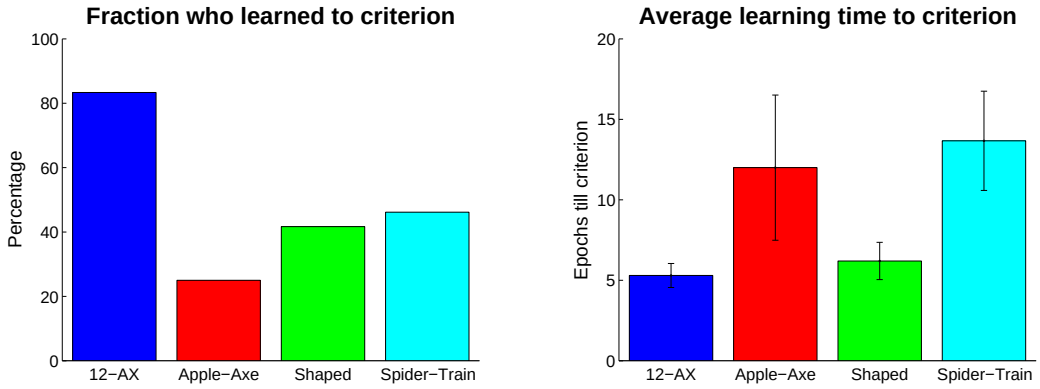
Since subjects progressed to the testing phase as soon as they had achieved the performance criterion, some of the training data for the full 25 epochs shown on the graphs was missing. In order to plot meaningful learning curves, subjects' data were filled-in for the remaining epochs after they had achieved the criterion or otherwise moved over to the testing phase early. This missing data was filled-in with each subject's average performance during the two testing epochs which followed training. As the testing trials were identical to training trials, apart from the neutral feedback, they were considered to reflect the subjects' likely continued training performance. If anything, performance should be worse during testing, as participants could not use the additional information of a signal to switch. Furthermore, correct responses could not be used by subjects to reassure themselves that they were in the context they thought they were.

In order to determine how well subjects learnt the task, a baseline performance was defined, akin to "chance level". This baseline was based upon the trivial strategy of always pressing the more frequently correct (i.e. default) button, independent of the stimulus seen. Thus, all non-target stimuli would be correct, as would on average half the potential target stimuli. According to the statistics, there were on average 2.5 inner loops per outer loop. For each outer loop, there was one context marker stimulus that was a non-target, and the first stimulus of the inner loop was also a non-target.

In 66% of cases, the second stimulus was a potential target and thus 33% were also non-targets. In the case of the experimental versions with a non-paired structure, probabilities for potential targets were also at 66% probability, but the distribution of non-targets was different. As the context markers were drawn uniformly, there was a 50% chance that a potential target was an actual target. Overall, this trivial strategy would lead to $\frac{1+2.5+2.5*(0.33+0.66/2)}{1+2*2.5} \approx 86\%$ of all responses to be correct. However, due to the weighting of non-targets vs. targets, the calculated reward would on average be $\frac{1+2.5+0.33*2.5}{1+2.5.+0.33*2.5+8.5*0.66*2.5} \approx 24\%$. Unlike in a true chance level, however, subjects can perform worse than this baseline level over sustained periods of time. Indeed during the first trials, where subjects did not know the asymmetry of the response, they did partially perform worse. It would obviously be unreasonable to compare to a strategy with equal probability for right and left responses, since this would lead to less than 0 reward on average.

In addition to the rejection of 3 subjects due to technical reasons (see 5.2.1), an outlier analysis was performed to detect any potential anomalies in subjects' data that should be excluded. The distribution of learning times per experimental condition of subjects achieving the criterion were analysed for outliers. All subjects' performance were within two standard deviations of the mean, giving no indication of outliers. Mean reaction times for subjects both during the learning as well as the assessment phase were also analysed and found to be all within 2 standard deviations of the mean. Results are thus presented only with all subjects included, as no subject met the criterion of an outlier. Although there was no statistical grounds to exclude subjects as outliers, by visual inspection there was one suspicious subject in the `12-AX` task. This subject achieved learning in only two epochs, with even the first epoch achieving more than 80% correct. Therefore, to verify that results were not influenced by a single potential outlier alone, all analyses were repeated with each subject excluded one at a time. It was then verified that this manipulation did not affect the significance of results. For the results involving only subjects who had learnt in the Apple-Axe condition this analysis was not possible as too few subjects were in this category.

In presenting graphs and figures for training the shaped Apple-Axe task, data generally do not include the training times of the shaping task themselves, unless stated otherwise. This allows us to compare the shaped and unshaped paradigms more directly. Training times on the shaping procedure are discussed separately in section 5.3.3.



(a) (b)

Figure 5.2: Learning performance of humans on variants of the 12-AX task: (a) Percentage of subjects who achieved the criterion and correctly learnt the full set of rules. (b) Average number of epochs subjects needed to achieve the criterion. Subjects who did not achieve criterion are excluded from this graph.

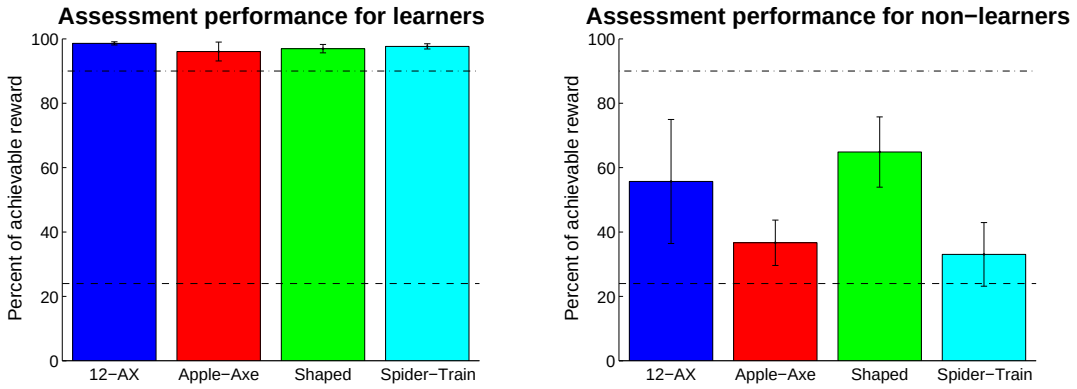
5.3 Results

5.3.1 Question 1: Can subjects learn the 12-AX task?

One of the main questions this study aimed to answer was, whether humans can rapidly learn the 12-AX task through trial and error, with only few hints in the instructions as to the structure of the learnt rules. It particularly focuses on whether the 12-AX task can be used in single session studies, to test performance under various conditions. For this purpose, subjects were tested on the base `12-AX` task and compared to the Apple-Axe variant.

As shown in figure 5.2a, on the simpler `12-AX` 83.3% (10 out of 12 subjects) achieved the learning criterion (2 consecutive epochs above 90% correct), and thus could be said to have learnt the full set of rules. In contrast, in the Apple-Axe version only 25.0% (3 out of 12 subjects) learnt the full set of rules, significantly less than in the 12-AX condition ($p < 0.01$, one tailed Fisher's exact test).

To verify the chosen learning criterion was appropriate, all subjects were tested in an assessment phase after learning. As shown in figure 5.3a, those subjects who achieved criterion during the learning phase performed well in the test, reaching an average of 98.6% of the achievable reward for subjects in the `12-AX` task and 96.0% for subjects of the Apple-Axe variant (difference not significant, two-tailed, two-sample t-test). Furthermore, all subjects achieved higher results than the 90% criterion (see figure A.5, A.7). During the test, subjects received no feedback and so could not use strategies such as win-stay, lose-switch to help with the outer loops. This further verified subjects' success in learning the full rules. In contrast, subjects who failed to reach criterion during learning gained less than 75% in both the `12-AX` and the Apple-Axe version, with



(a) (b)

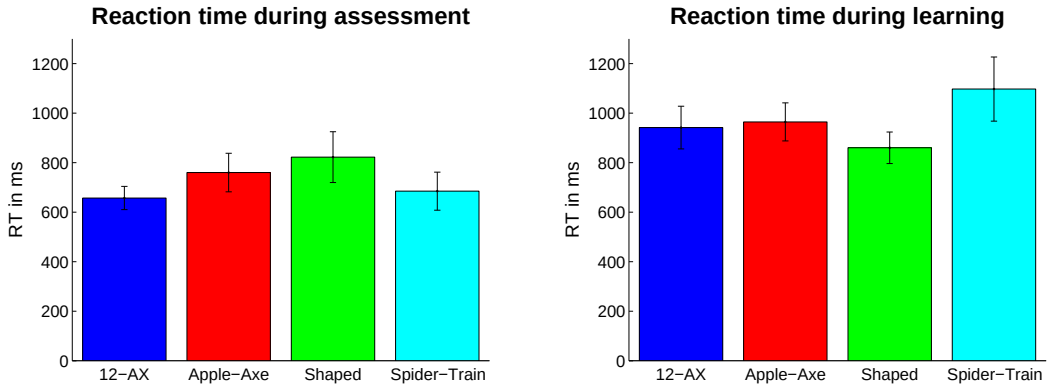
Figure 5.3: Percentage of achievable reward during the assessment phase after training broken up by learners and non-learners: (a) Subjects who have learnt the task to criterion performed well during assessment and made few errors only (b) Shows performance during assessment for subjects who did not achieve the required criterion during training.

the means being 55% and 37% respectively as shown in figure 5.3b (again, difference not significant, two-tailed two-sample t-test).

As an additional confirmation of the criterion, subjects were asked to explain the rules they had identified during a debriefing session after the end of the experiment. All but one subject (performing the Spider-Train version), achieving the learning criterion, could explain the rules, albeit sometimes with substantial difficulty of verbalising them in a coherent manor (see 5.4 for more details on the debriefing). In contrast, no subject who had not achieved criterion during either training was able to identify the full set of rules, although several sub-components of the task were correctly explained to varying degrees.

How quickly do subjects learn?

Amongst those subjects who learnt, the average numbers of epochs taken were 5.3 ± 0.7 for the `12-AX` version and 12.0 ± 4.5 for the Apple-Axe task (figure 5.2b). With an average number of approximately 150 stimuli per epoch (148.1 ± 2.0, 149.8 ± 0.8 for `12-AX` and Apple-Axe respectively), this corresponded to subjects learning the `12-AX` task in approximately 795 stimulus presentations on average and the Apple-Axe in 1800. Therefore, not only were fewer subjects able to learn the Apple-Axe task, but those that did, also needed significantly more epochs to do so ($p < 0.05$, two-tailed two-sample t-test).



(a) (b)

Figure 5.4: Overall reaction times independent of learning status: These graphs show the mean reaction times averaged across all subjects, independent of their learning result. Averages are not corrected for outliers. (a) Averages for epochs in the assessment phase (b) Averages for epochs throughout the learning phase.

Basic reaction times in light of self-timing

The subjects were self-timed with up to 20 seconds to respond per stimulus. In the assessment phase, after subjects had completed learning, the average reaction times, averaged over all subjects, were $657ms \pm 47$ and $760ms \pm 78$ for the 12-AX and the Apple-Axe versions respectively, as shown in figure 5.4a (the difference did not reach significance $p > 0.1$, two-tailed two-sample t-test). Thus on average, reaction times, despite being self-timed, were below 1 second. However, the per-subject average RT varied considerably with a range of $393ms - 900ms$ and $337ms - 1316ms$ for the two conditions. Whether subjects had learnt the task did not significantly effect the RT during the assessment phase, as is evident in table 5.1.

During the learning phase, reaction times were overall unsurprisingly slower, as shown in figure 5.4b. However, averaged over all conditions, including the shaped and Apple-Axe conditions, subjects still responded to about 1 stimulus per second ($970ms \pm 48$) during training. Again, per subject averaged RTs (calculated over the complete training phase) were not significantly different between the versions (figure 5.4b). Differences in mean reaction times during the learning phase between subjects, who learnt the task and those who didn't, were also not significantly different (table 5.1). The unconstrained timing did, however, result in a heavy-tailed distribution of reaction times. Some responses took several seconds, resulting in an average kurtosis of the per-subject RT distributions during learning in excess of 20.

Despite the lack of any overall differences, breaking the RTs down by sequence and state revealed interesting differences, allowing us to guess at some of the strategies they adopted. Figure 5.7 presents relative reaction times comparing (unconditional) responses to the potential target stimuli X or Y and all other stimuli never requiring

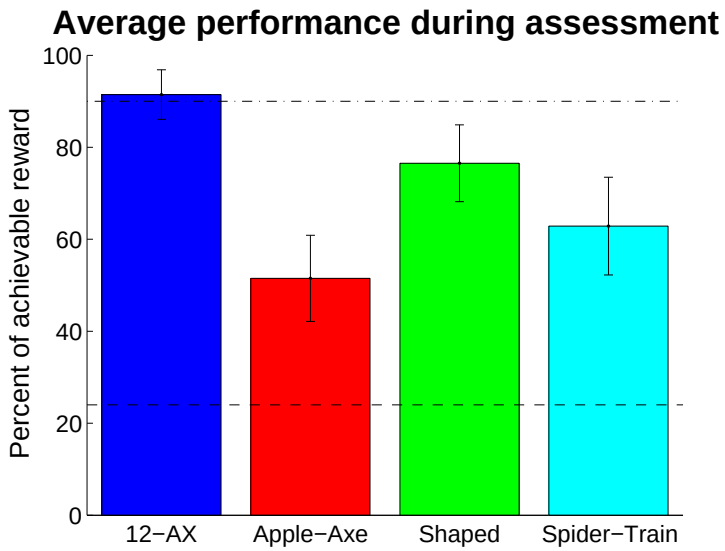


Figure 5.5: Average performance in the assessment phase: This graph shows subjects' performance in the assessment phase after undergoing training. Performance is shown as the percentage of achievable reward, where the reward is calculated according to the asymmetric loss function. Results are averaged over all subjects. The lower dashed line at 24% represents "chance level", whereas the upper denotes the learning criterion.

a target response. As shown in figure 5.7a, amongst subjects who learnt the rules, there was no significant difference in reaction times between potential targets and non-targets for all but the `12-AX` version. The latter in fact showed faster reaction times for potential targets than non-targets ($p < 0.01$, two-tailed two-sample t-test, per subject RT variations were normalised via z-scores). This was in contrast to subjects in the non-learner group (figure 5.7b) who, for all but the shaped Apple-Axe version, showed highly significantly ($p < 0.001$) slower reaction times to potential targets than non-targets. This dissociation was expected for the following reasons: For all sequences ending on a potential target but the AX and AY in the 1 context, and BY and BX in the 2 context, the correct outcome is already known by the previous stimulus. For these 4 cases, however, the strong difference in prevalence of AX and BY over AY and BX creates a form of default response, allowing to predict the correct response to the vast majority of X and Y stimuli. For subjects who do know the rules, stimuli X and Y have limited uncertainty and need no context switch resulting in shorter RTs. For subjects who don't know the full rules on the other hand, the potential targets are the only stimuli that have uncertainty as to how to react, hence one would expect the slower RTs to these stimuli.

Intermediate conclusion

These results thus confirmed that the base `12-AX` task could be learnt without major problems under these conditions. In contrast, significantly fewer subjects (posterior

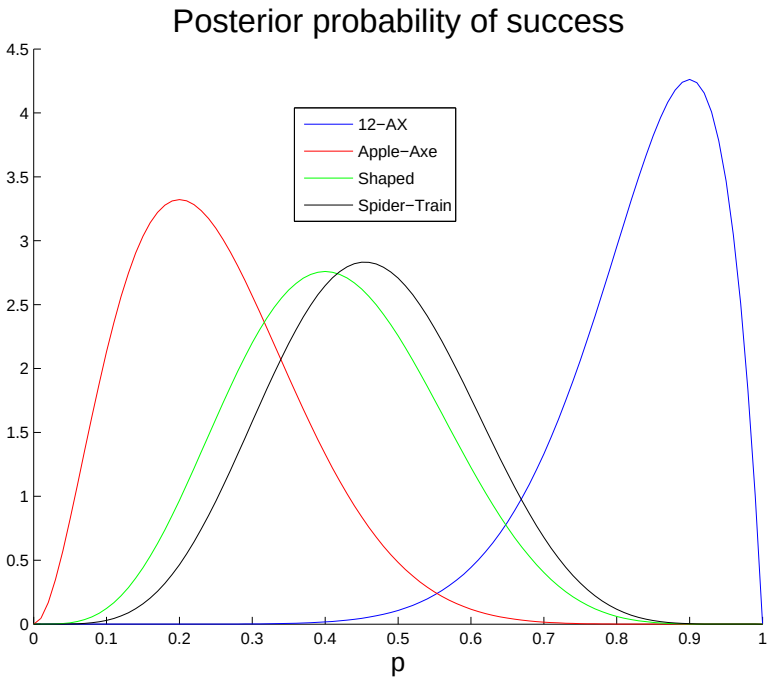


Figure 5.6: This graph shows the posterior distribution of the probability of successfully learning each of the task conditions, based on a flat beta prior.

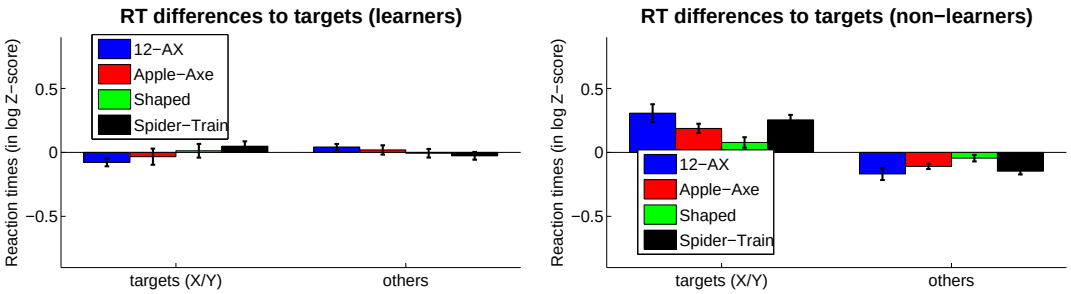
distribution of the learning probability is shown in figure 5.6) were able to learn the Apple-Axe version. Thus as expected, the Apple-Axe version was more difficult to learn. These results also show that the different conditions cover a wide spectrum from quickly (easily) learnt by the majority to only learnable by a few over a longer time during a single session, all within the same task structure.

5.3.2 Question 2: Is the categorisation of stimuli a possible explanation of increased performance?

One possible reason for the large differences in performance between the `12-AX` and the Apple-Axe version is that the stimuli in the `12-AX` task are categorised according to the

	`12-AX`	Apple-Axe	Shaped	Spider-Train
RT during assessment				
all subjects	$657ms \pm 47$	$760ms \pm 78$	$822ms \pm 102$	$684ms \pm 77$
learnt	$654ms \pm 54$	$807ms \pm 155$	$689ms \pm 52$	$602ms \pm 66$
not-learnt	$674ms \pm 126$	$745ms \pm 95$	$899ms \pm 155$	$756ms \pm 130$
RT during learning				
all subjects	$942ms \pm 86$	$965ms \pm 77$	$860ms \pm 63$	$1097ms \pm 129$
learnt	$973ms \pm 97$	$759ms \pm 83$	$856ms \pm 87$	$948ms \pm 114$
not-learnt	$786ms \pm 186$	$1033ms \pm 89$	$863ms \pm 91$	$1226ms \pm 217$

Table 5.1: This table shows the mean RTs for the various conditions



(a) (b)

Figure 5.7: Relative reaction times between non- / targets: These graphs show the relative reaction times in response to different stimuli during the assessment phase. They compare response times to potential target stimuli (`X` or `Y`) to an average of all other stimuli. (a) shows the average for subjects who have learnt the task and (b) for those who haven't.

task structure (context markers are numbers; `A` and `B` occur early in the alphabet; `X` and `Y` occur late). This provides additional information about the task. The Spider-Train version of the task was used to test this. Keeping all other parameters the same as in the Apple-Axe version (stimuli from the same overall set, equal unpaired structure), the stimuli (shown in figure A.3) were chosen such that their categorisation aligned with the task structure. That is, context markers were chosen from the category of insects, target stimuli from the category of vehicles and inner-loop markers were chosen from amongst the musical instruments. With this variation, 46.0% learnt the rules to criterion (6 out of 13 subjects) (see figure 5.2a). This version of the task was expected to be quicker to learn than the Apple-Axe task, but harder than the `12-AX`. Nevertheless, although the fraction of subjects who learnt lay in-between that for the two other tasks, the statistical power was insufficient to determine a significant difference between either (Fisher's exact test), although a trend existed for the comparison with the `12-AX`. Therefore, these results remain inconclusive with respect to if and how much of the difference between the Apple-Axe and `12-AX` versions can be attributed to the categorisation of the stimuli. The Spider-Train took an average of 13.6 ± 3.0 epochs to learn. This is not significantly different from the Apple-Axe, but like the Apple-Axe, is significantly different from `12-AX`

One potential compromising factor is the difference in difficulty in processing the stimuli, which might be revealed through the RTs. However, comparing the number and letter stimuli (`12-AX` version) vs the abstract pictorial stimuli (Apple-Axe, Spider-Train and shaped Apple-Axe averaged together) for learners, the difference in RT ($654ms \pm 54$ vs $676ms \pm 50$) was not significant. Comparing RTs for the Spider-Train version to any of the three other versions individually also did not result in any significant differences; neither for RTs during learning or assessment averaged over all subjects (figure 5.4a, b), nor split according to non- / learners. (table 5.1). Thus, RTs did not provide evidence for the difference between the `12-AX` and the Apple-Axe version originating from the

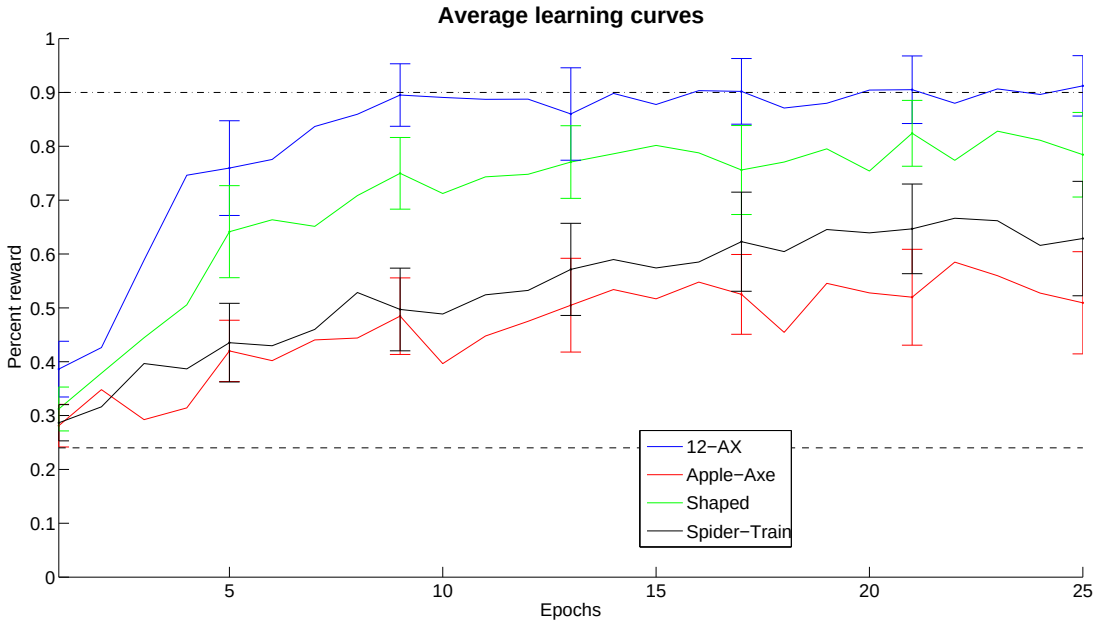


Figure 5.8: Learning curves for the 4 experimental variants: The learning curves are averaged across all subjects in each of the conditions. For subjects completing the training session before epoch 25, their data was extended with their respective assessment performance. The graph in all cases shows training during the complete task and does not include training of the shaping tasks. The dashed lines indicate the learning criterion at 90% and the expected reward under the trivial strategy of 24%. Dotted lines represent one standard error of the mean.

difference in difficulty of their respective stimuli.

5.3.3 Question 3: Do humans benefit from shaping in the 12-AX?

The third main question was whether humans could take advantage of a shaping protocol akin to that used in modelling in chapter 4. After undergoing the shaping procedure described in 5.2.3, 41.7% of subjects (5 out of 12, two additional subjects were excluded for not completing the full experiment) learnt the full 12-AX rule set (see figure 5.2a). The proportion of subjects able to learn to criterion was not significantly higher than that of the Apple-Axe version. However, as shown in figure 5.5, in the assessment phase of the shaped Apple-Axe condition, subjects achieved on average a test performance of $76.5\% \pm 8.3$ of achievable reward, compared to an average of only $51.5\% \pm 9.3$ for the plain Apple-Axe version. This provided a trend ($p < 0.07$) towards the effectiveness of shaping. Indeed, when looking at the full learning curve in figure 5.8, this trend was confirmed. Shaped subjects do learn significantly more competently than unshaped subjects in the Apple-Axe condition (significant from epoch 9 onwards).

Amongst those subjects who did learn in the shaped Apple-Axe condition, the average number of epochs needed was only 6.2 ± 1.2, close to the 5.3 epochs needed by subjects of the `12-AX` version. This can again be seen in the learning curve of those subjects who achieved the criterion (figure 5.9). Learning for the `12-AX` and the shaped Apple-

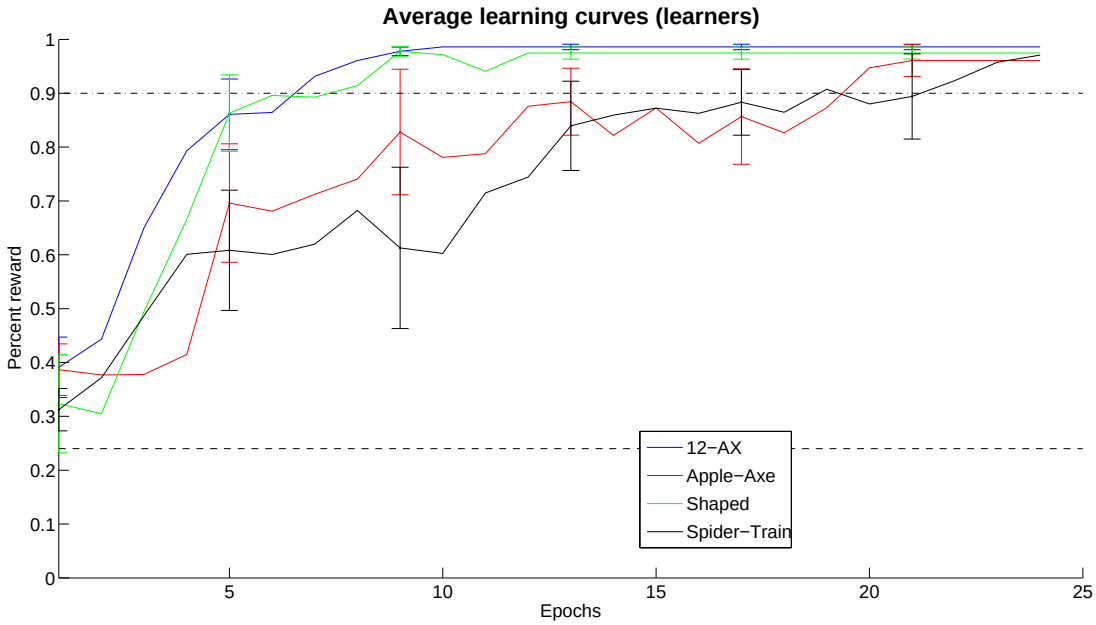


Figure 5.9: Learning curves for those subjects successfully achieving criterion.

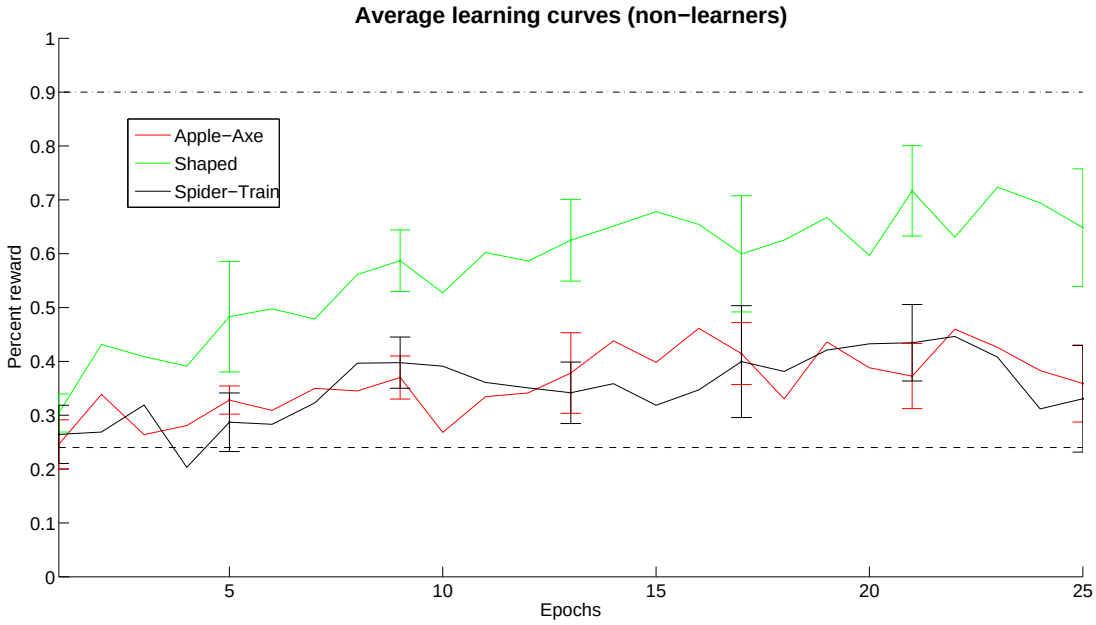


Figure 5.10: Learning curves for those subjects who did not achieve criterion.

Axe versions rose to criterion faster than either the plain Apple-Axe or Spider-Train versions.

An even bigger difference, though, was evident in the performance of subjects who did not achieve criterion. Non-learners in all versions, as shown in figure 5.3b, performed better than the trivial strategy of pressing non-target to all stimuli (which results in 24%). Only for the shaped Apple-Axe, though, was this highly significant ($p < 0.005$, one-tailed t-test), with a trend ($p < 0.06$) for the Apple-Axe version (there were only 2 non-learners in the `12-AX` condition, making it impossible to test the hypothesis for this condition). Further, performance during assessment of shaped non-learners was significantly higher ($p < 0.05$, two-tailed two-sample t-test) than for subjects of the non-shaped Apple-Axe and Spider-Train versions ($64.8\% \pm 11$ vs $35.1\% \pm 5.7$).

Further evidence towards the higher performance of shaped non-learners was provided by the RT difference analysis of potential targets vs non-targets. As shown in figure 5.7b, slowing to potential targets was less pronounced in the shaped condition, and not significantly different to the RT differences of the learner group, unlike those of the other versions.

Learning the shaping tasks

As intended, all three of the individual shaping tasks were learnt more rapidly than the full 12-AX task structure. Together, the 3 shaping tasks were mastered in a average of 11.3 ± 1.5 epochs, split reasonably evenly between the tasks: 4.1 ± 0.6, 3.4 ± 0.6 and 3.8 ± 1.7 for the 12-task, the AX-task and BY-task respectively (see figure 5.11). Indeed, all subjects successfully achieved criterion for each of the shaping tasks. Apart from one outlier subject, who took 19 epochs to learn the BY task after needing just two epochs each for the preceding 12 and AX tasks, all shaping tasks were completed within a maximum of 8 epochs; several were learnt within the first or second epoch. In a comparison between subjects who successfully learnt the final task and those who didn't, there was no significant difference in learning times through the shaping sequence.

5.3.4 Question 4: Do non-learners partially pick up on substructure of the task?

In order to try and understand further the degree to which non-learners of the full task learnt at least sub-parts of the rules, the results were decomposed during the assessment phase. It was first checked whether subjects realised that only certain sub-sequences could ever be part of the target sequence. Subjects had no trouble identifying that only stimuli X and Y (or their equivalent) ever required a "target response". Figure 5.12b shows the proportion of correct responses to stimuli that subjects had responded to as

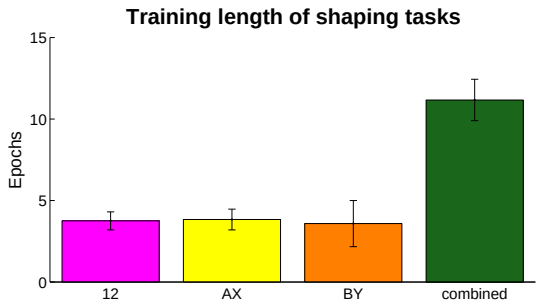
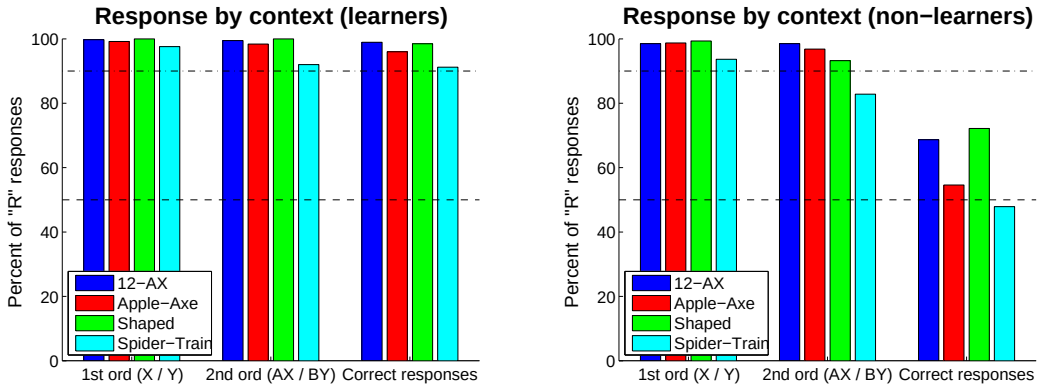


Figure 5.11: Number of epochs spent in each of the three shaping stages: This graph shows the average number of epochs subjects spent in the three stages of shaping. The criterion used to judge when to progress to the next stage was one epoch above 90% correct. The "combined" bar represents the average time subjects spent in all three tasks together before entering the full Apple-Axe task. The graph includes all subjects, independent of their later success with the full Apple-Axe task.

targets. These are separated into three groups according to the "sequences" of varying lengths in which they were embedded. The three groups split into first and second order sequences as well as full correct responses. First order refers to sequences of length one, i.e. the stimulus alone; second order refers to inner loop sequences AX and BY and the full group to AX and BY sequences in their respective correct context. On this first-order statistic non-learners of all tasks performed well during assessment (98.5%, 98.7% 99.4 and 93.7%) and showed no significant differences to subjects who had achieved criterion (98.8%, 100.00% 98.5%, 95.0% see figure 5.12a). For all but the Spider-Train version, subjects also correctly identified the second-order dependency, i.e. that only sequences of AX or BY could be targets. Further, for all but one subject in these conditions (the exception being in the shaped Apple-Axe condition) more than 92% of their target responses followed a potential target sequence. For the Spider-Train, though, the average was only 82.8%, with 2 subjects achieving 100% and the other 5 out of 7 subjects achieving less than the 90%. The lowest was 67%.

The next layer of structure concerns the switching nature of the context markers, asking the question if subjects correctly identify their persistence across multiple inner loop sequences. The percentage of targets to which subjects correctly responded depending on the distance between the inner loop (AX and BY) and the context markers (1 and 2) (figure 5.13b) was assessed. If subjects indeed only paid attention to the immediate sequences and did not take the switching behaviour into account, the performance for context markers immediately preceding the inner loop should be higher than those with intermediate inner loops or distractors. The data are consistent with this, in that for all 4 conditions, the immediate sequence targets have a higher percentage of correct responses. However, these differences individually are not significant. Furthermore, despite the explicit training of the switching behaviour in the first task of shaping, there was not a strong difference according to this metric between the shaped Apple-Axe version and the standard Apple-Axe condition, other than the overall higher performance



(a) (b)

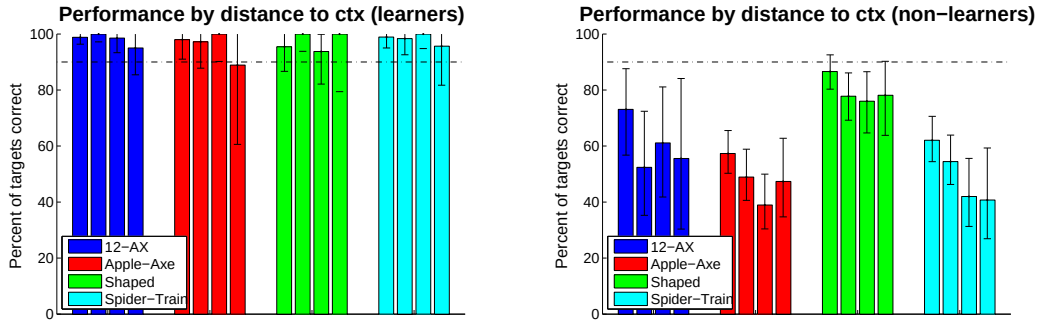
Figure 5.12: Target responses during the assessment phase, broken down by context: The first bar-group (1st order) is the percentage of "target" responses that were made in response to one of the two possible target stimuli (`X` / `Y`). The second order bar-group is the percentage of target responses that were in response to one of the two target sequences `AX` or `BY`. The third group (correct responses) is the percent of target responses that were correct under the full set of rules. As all are conditioned on the subject responding with the target button, errors shown here are only *commission errors*, not omission errors. Colours are as in the other plots. (a) Responses conditioned on the subjects having achieved criterion and thus considered having learnt the task. (b) Responses for non-learners.

in the shaped case.

5.3.5 Question 5: Are there different types of learners?

As expected in a trial and error style learning of a complex task, there is a significant amount of variance both in number of epochs required to achieve learning, as well as in the epoch to epoch error rate. One possible source of such variance would be the existence of different learning strategies across subjects. To try and identify whether there are different groups of learners, potentially hinting towards different strategies, the individual learning curves of subjects were compared. One indication of distinct types of learners would be a multi modal distribution of learning curves. In order to have a higher number of samples, learning curves of all 4 conditions (`12-AX`, Apple-Axe, Spider-Train and shaped) were combined in the same graph. As shown in figure 5.14 there is a compact cluster of subjects achieving criterion within 12 epochs. In addition there are three further subjects who achieve criterion later. However, their learning curves are not sufficiently distinct to determine if they are part of a heavy tailed distribution or form a distinct cluster hinting towards a separate strategy.

Amongst those learners who required more epochs to achieve criterion, individual learning curves show a non gradual learning pattern, i.e. an initial phase resembling that of the non-learners, followed by a rapid increase to criterion.



(a) (b)

Figure 5.13: Target responses by context depth: These graphs show the percentage of correct target responses made by subjects during the assessment phase, broken down by their distance to the last context marker (1 / 2). The first column shows the percentage correct for target sequences (AX / BY) immediately following a context marker with no intervening stimuli. The second, third and fourth columns show target responses with one, two or three inner loops between the respective target sequence and its corresponding context marker. For non-paired paradigms, responses of both positions are combined, i.e. column 2 contains sequences with either one or two intermittent stimuli. (a) Responses conditioned on the subjects having achieved criterion and thus considered having learnt the task. (b) Responses for non-learners.

One hypothesis for distinct strategies of learning is a hypothesis-testing approach vs an associative learning style. Due to the computational properties of the two strategies, a differential efficiency in the use of samples would be expected. Furthermore, one might postulate that both strategies differentiate themselves based on reaction times, with the hypothesis-driven approach being the slower of the two. This would suggest, that subjects with a longer RT during learning would be faster in achieving the learning criterion. To test this prediction, the number of epochs taken to achieve learning was correlated with the mean RT during training. Although a negative correlation exists in the data, it is weak ($R^2 = 0.139$) and not statistically significant.

Looking at subjects who did not achieve criterion a similar picture emerges (figure 5.18). These subjects show a homogeneous distribution of per epoch performance throughout the training session.

Within the available data, there is overall no conclusive indication of categorically different types of learners or non-learners that would allow these groups to be distinguished further. In contrast, the groups of learners vs non-learners do show distinct learning curve distributions, as shown in figure 5.17. This indicates that the split into learners vs non-learners is not simply an artifact of the cut-off criterion at 25 epochs splitting a uniform distribution in two. Amongst learners, the cut-off criterion is at 2.5

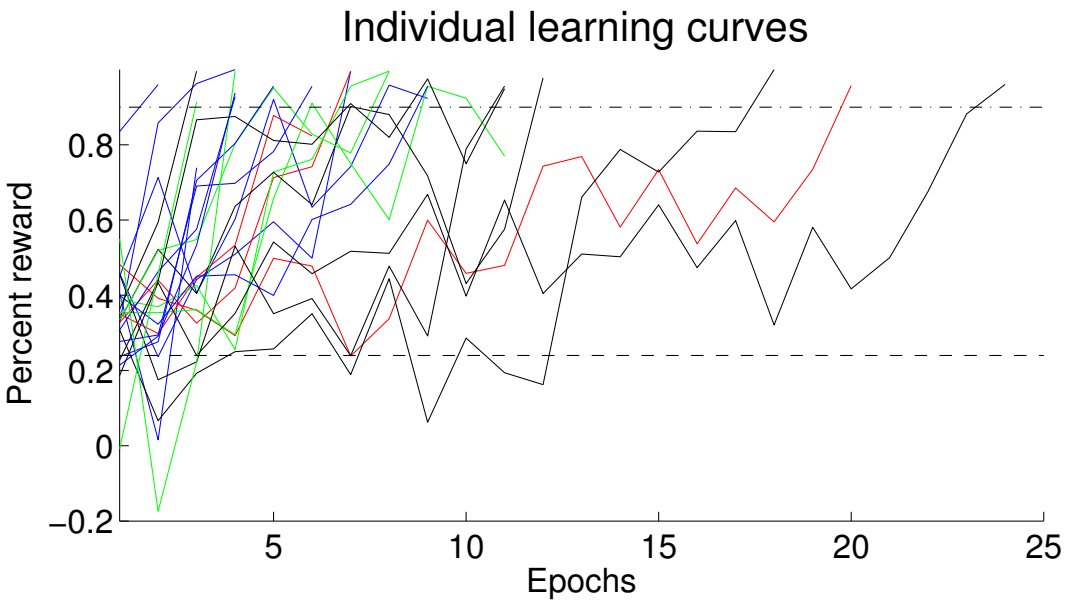
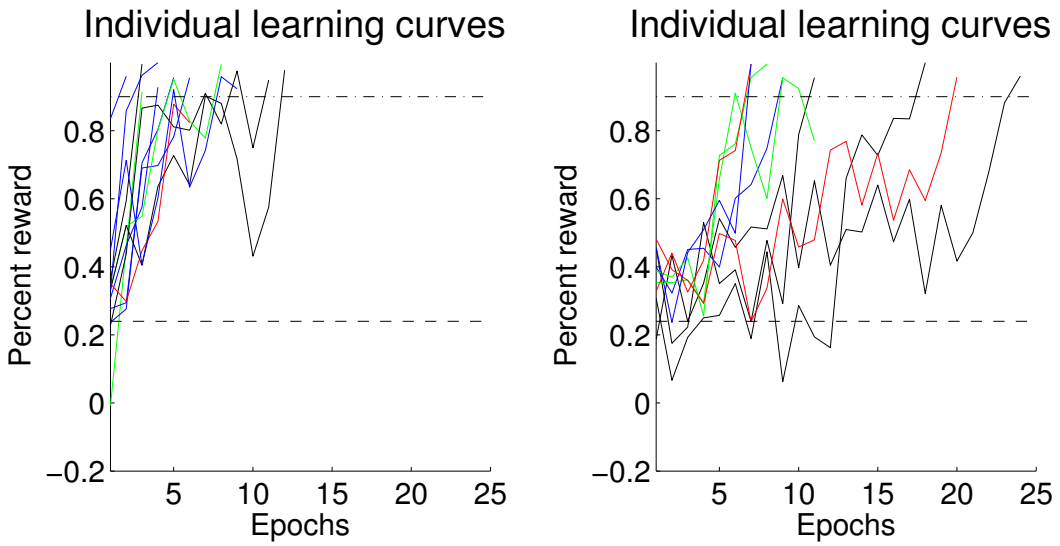


Figure 5.14: Individual learning curves for learners: This graph show the learning curves of individual subjects who have achieved the learning criterion. Blue are subjects in the `12-AX` condition red are subjects in the Apple-Axe condition, green are in the shaped condition and black in the Spider-Train condition



(a) (b)

Figure 5.15: Individual learning curves for learners: These graphs show the learning curves of individual subjects who have achieved the learning criterion. They are split according to a visual examination into a group of subjects who show a near monotonic increase from the start and one which shows an initial pattern of non-learner before rapidly reaching criterion. Blue are subjects in the `12-AX` condition red are subjects in the Apple-Axe condition, green are in the shaped condition and black in the Spider-Train condition

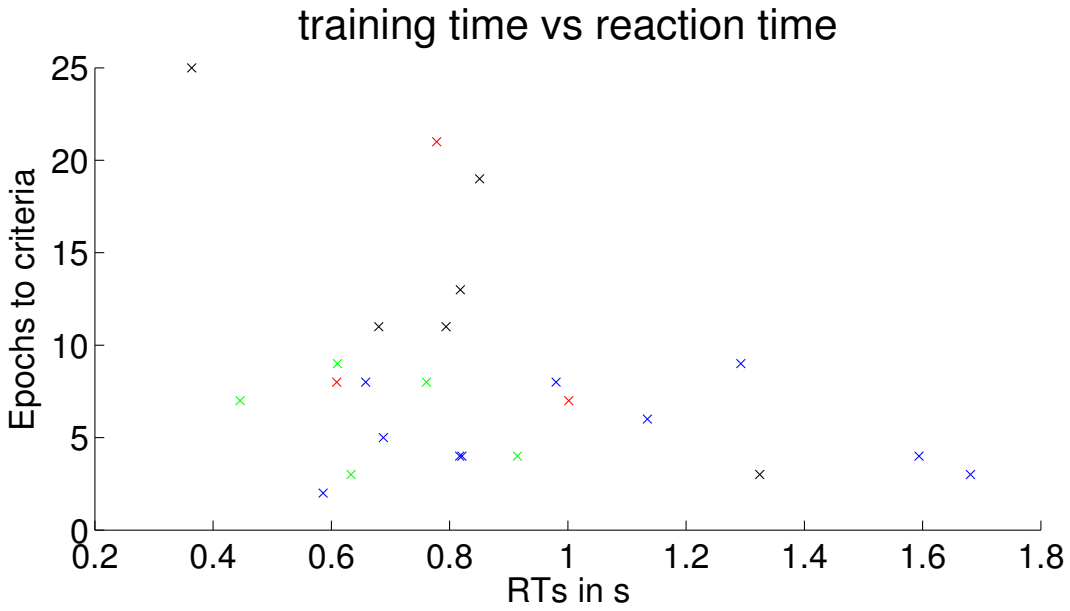


Figure 5.16: Correlation between RTs and learning time: This graph shows the relation between the average per subject reaction time during the training phase and the time it took them to reach the learning criterion. Only subjects who have achieved the criterion are included. Blue are `12-AX`subjects, red are Apple-Axe, green are shaped and black are Spider-Train

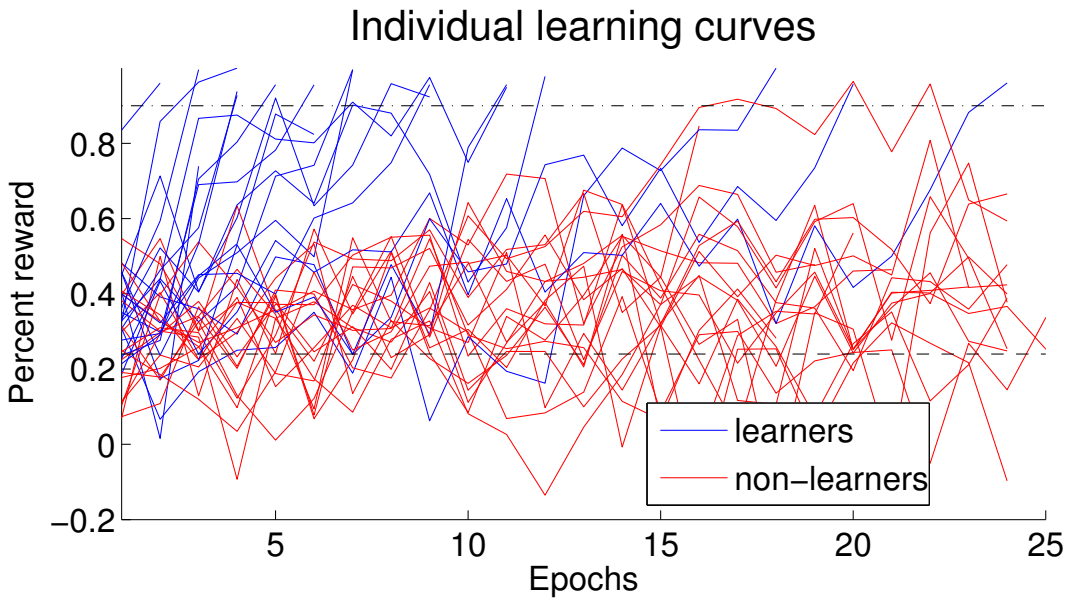


Figure 5.17: Individual curves of learners vs non-learners: This graph contrasts individual subjects learning curves between the group of learners and non learners. It includes subjects of the `12-AX`the Apple-Axe and the Spider-Train condition. It shows the mostly very distinct patterns of learners vs non-learners.

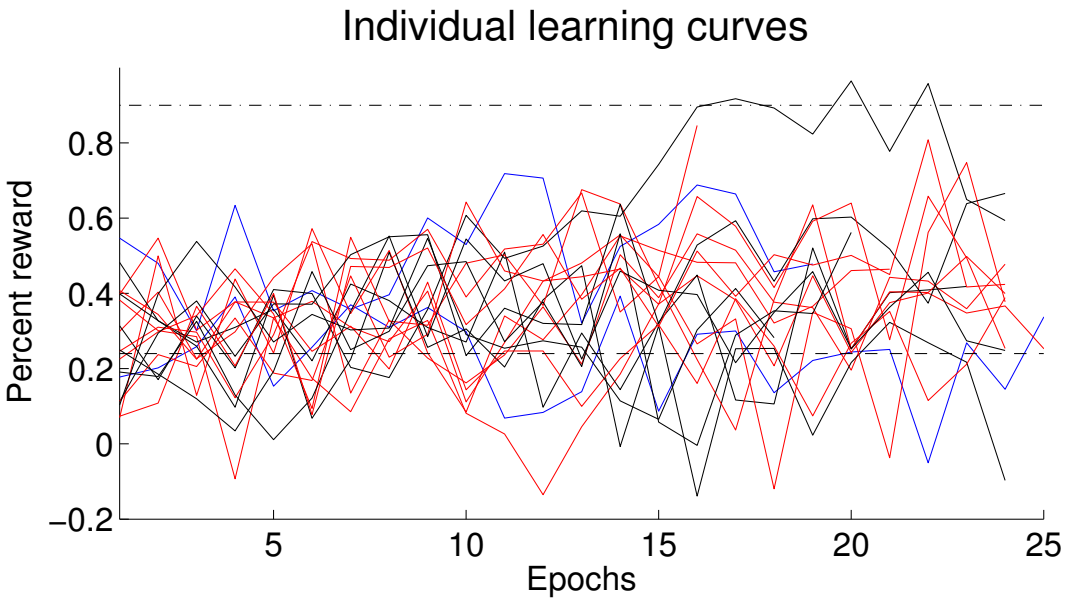
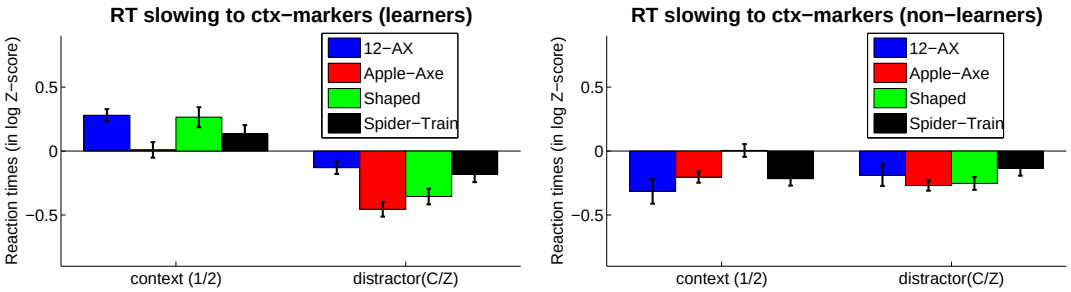


Figure 5.18: Individual learning curves for non-learners: These graphs show the learning curves of individual subjects who have not achieved the learning criterion. Blue are subjects in the `12-AX` condition red are subjects in the Apple-Axe condition, green are in the shaped condition and black in the Spider-Train condition



(a) (b)

Figure 5.19: Reaction times to context markers: These graphs show the relative reaction times in response to different stimuli during the assessment phase, comparing response times to the context marker stimuli (`1` or `2`) to the average of all other stimuli. (a) shows the average RT for subjects who had learnt the task and (b) shows them for those who hadn't.

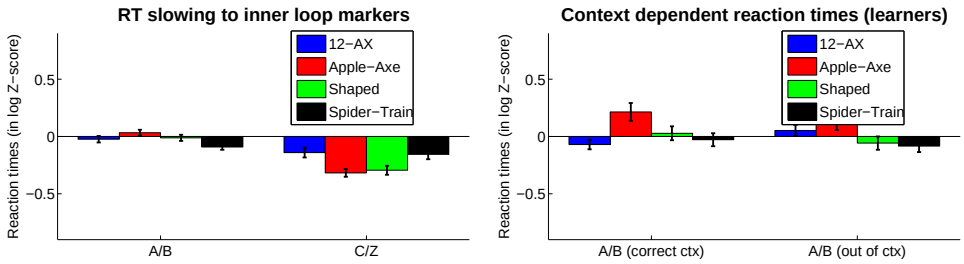
standard deviations above the mean.

5.3.6 Gating into working memory, a reaction time analysis

An important concept in the models of previous chapters was gating, i.e. the act of storing percepts into working memory. RTs were analysed to look for evidence of such a process. From the models, one would predict, that subjects who learnt the rules would show slower reaction times to the context stimuli than other stimuli, as these require a "gating action" or "task switching". Indeed, when comparing response times to context markers to the distractor stimuli C and Z, that have no task meaning (see figure 5.19a), all versions showed a significant and large slowing to the context markers. This is in strong contrast to subjects who had not learnt the rules to completion. Further, the differences between context markers and distractors disappeared for all but the shaped Apple-Axe. This supported the hypothesis that non-learners did not recognise the context markers as being relevant, and did not gate the information or switch the task, as suggested by the choice data of figure 5.12b. When comparing the shaped Apple-Axe version to all non-shaped versions, non-learners no longer showed the faster reaction times of distractors for the context markers, but instead were significantly slower. This shows that with shaping, even the non-learners reacted to the context markers as trained, on remembering the 1 and 2 during the first stage of shaping.

Further, in the models, gating and storing of A and B stimuli are treated as being equivalent to the gating and storing of the context markers 1 and 2. These just occur one level down the hierarchy. As RTs were significantly slower for the 1 / 2 context, one would predict a similar slowing would occur to the A and B stimuli, which also need remembering. These responses could be averaged over all subjects, learners and non-learners, given even the latter were able to cope with these sequences. In doing so, it was found that for all conditions but the Spider-Train version, responses to the A and B stimuli were slower than for the distractors (figure 5.20a). The effect however, was not as big as between context and distractor stimuli. Looking to see if there were differences between RTs, when A and B appeared in their respective correct context vs incorrect context, did not reveal a significant difference; either for learners alone (figure 5.20b), or for the full group.

Apart from gating, RT differences to the context markers might also be explained in terms of task switching costs. To investigate this further, the differences in relative reaction times to context markers that either switched the context or kept it the same were looked at. For subjects who learnt, RTs were significantly higher to switching context markers than to those repeating the context in the conditions of the `12-AX` and shaped Apple-Axe (two-tailed, paired t-test on z-scored RTs). For subjects in the Apple-Axe or Spider-Train task this effect was not significant. Similarly, for subjects who did not learn, this difference was not present in any of the conditions. Therefore, at least under some conditions, there is a clear switching cost involved. However,



(a) (b)

Figure 5.20: Reaction times to inner loop gating: These graphs show the relative reaction time differences to the inner loop stimuli. (a) shows the reaction times to inner loop markers A and B compared to distractor stimuli C and Z, averaged across all subjects independent of learning condition. (b) shows reaction times for inner loop stimuli depending on context, i.e. either response to A in the 1 context and B in the 2 context or vice versa. Data is averaged across subjects who have learnt.

when RTs to non-switching context markers were compared between learners and non-learners, the non-switching context markers of learners still showed a significant slowing, indicating that switching is not the only cost associated.

Predictability of responses in the learnt task

Finally, when looking at more complete sequences that allow to fully determine a target or non-target response, similar patterns emerge, as shown in figure 5.21. Contingent on subjects having learnt the `12-AX` task (see figure 5.21a 5.21c), those sequences, that are either fully predictable after the stimulus prior to the measured one (`BY` in the `1` context and `AX` in the `2` context) or which follow the more likely outcome (`AX` in the `1` context and `BY` in the `2` context), showed significantly faster reaction times than average to the `X` or `Y` stimulus in the sequence. These faster reaction times did not depend on the the actual physical response, as it was shown by both target response (`1AX`, `2BY`) and non-target response sequences (`1BY`, `2AX`). In contrast, responses to the conditionally less likely outcome of non-target after seeing an `A` in the `1` context or a `B` in the `2` context were slower. Again, when comparing these results to subjects who had not learnt the task, these differences disappeared and all 3 sequence combinations showed the slowing expected from the marginal reaction times in figure 5.7b.

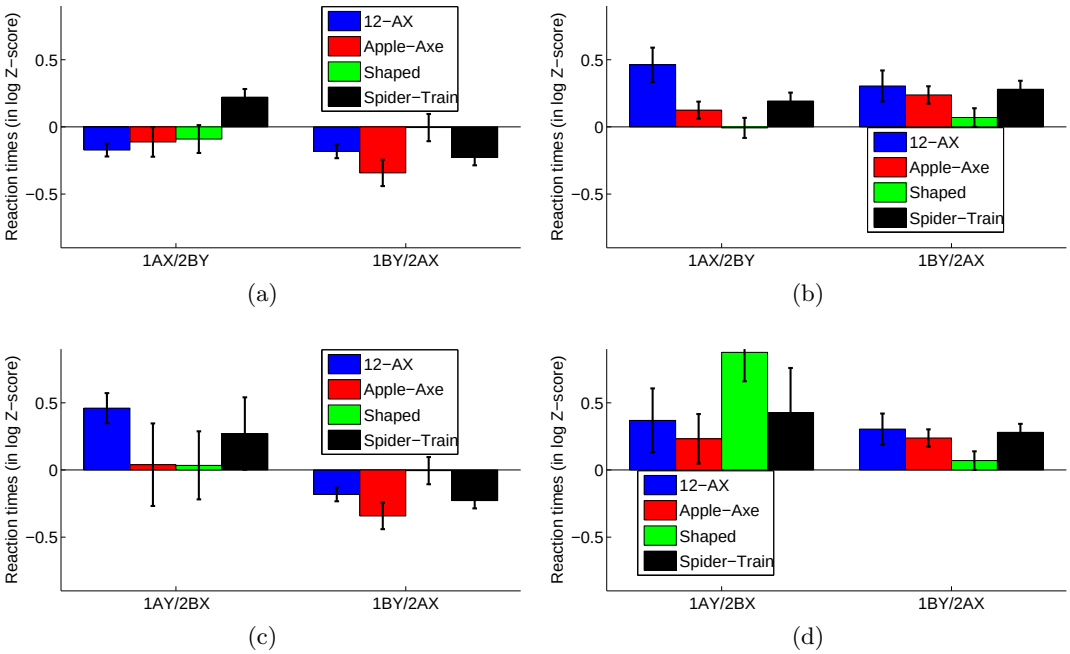


Figure 5.21: Reaction times to anticipatory responses: These graphs compare reaction times between responses that don't match the more common pattern and those that can be predicted by the structure of the task at least one stimulus prior to the one measured. In all cases, the RTs were measured for the same set of stimuli of potential targets (X and Y). (a)-(b) compare target responses vs non responses. Both are however by default predictable at the time of the measured X and Y. (c)-(d) compare reaction times for anticipatory and non anticipatory non-responses. (a) and (c) include subjects who have learnt to criterion, (b) and (d) are the groups of non-learners.

5.4 Debrief

At the end of each experiment, subjects were asked to describe in words the rules they thought governed the responses in the experiment. To aid their explanation, subjects were given a copy of all 8 stimuli to remind them and to allow them to point out the rules. However, as the debrief was verbal and unstructured, subjects reported both the rules and their observations about how they discovered these rules in very different ways. Thus subjective interpretation of the debrief was necessary and no rigorous quantitative analyses was performed. Nevertheless, it provided helpful validation of the quantitative analysis and provided insight into various strategies attempted by subjects.

All subjects in the `12-AX` version who achieved learning criteria were able to describe the rules correctly. Descriptions were good and succinct without much difficulty in explanation in all but 3 subjects, who gave more elaborate and convoluted explanations. Subjects did not express any frustration with the task, with 3 actually describing it as "fun". One expressed during the debrief that they found the task "too easy". In the other three conditions, all but one subject who had learnt to criterion, was still able to describe the task. However, subjects explained it less eloquently and with more difficulties than those in the 12-AX task. While all subjects in the `12-AX` condition explained the rules without referring to the stimuli sheet, subjects in the other conditions did at least look at the copy of the stimuli sheet as a memory aid. 5 subjects even directly used the copy to point to the stimuli while explaining the rules.

The debriefs for subjects who did not achieve criterion were in line with quantitative analyses presented above. However, given the variability of what subjects reported, no reliable differences between versions of the tasks were identified. As such numbers are reported across conditions and are predominantly to indicate potential strategies used. All but 2 subjects could correctly identify that the target was only either the (equivalent of the) X or the Y stimulus. 8 subjects correctly identified that the context marker stimuli had some role to play, in that the AX and BY were not always rewarded, but had not realised the switching between the two target sequences. Another subject identified the switching behaviour on 2 and realised that the 1 is relevant, but did not correctly connect them to the full rules. 5 subjects suggested that the context markers were part of a longer sequence of 1AX and 2BY. One participant went even further and tried to explain the full task in terms of ever more elaborate combinations of sequences. This enabled him to achieve a fairly high accuracy, since those combinations that did not fit into his schema were rather rare, due to the random sampling. Using this method, he achieved individual epochs above 90%, but not the more stringent criterion of two consecutive epochs above 90%.

2 other subjects tried to construct elaborate explanations involving the identity of the stimuli to explain the task in form of a story like "the hand is used to play ball, so those belong in a sequence, and the air plane carries away the person (hand) and so switches

away the sequence". How important such explanations were to learning is unclear, and remains an important question whose implications are discussed further below.

During the debrief subjects indeed provided anecdotal evidence for both explicit and implicit learning. Fore example, one subject achieved both the learning criterion during training and a high score during assessment, but afterwards explained a different task and even claimed after being told the full rules that this was not what they were doing. Other subjects showed signs hinting towards more explicit learning. One subject even uttered a satisfied "Ah-Ha" during training, after which he performed the task correctly.

In the shaping condition, 2 subjects mentioned the relation between shaping stages and the final task correctly. A further 3 subjects identified the relation correctly after being asked if they noticed a relation. Other subjects, however, were not able to correctly report this even after explicitly being asked. 5 out of 12 subjects were not able to fully recollect the rules of the three shaping stages, even though reaching the learning criterion.

5.5 Discussion and conclusion

The 12-AX task was originally designed to highlight specific needs and challenges for computational learning architectures, particularly selective gating. It has since featured in several different studies of learning. However until now, no data on how humans learn this task existed. Such data are necessary to enable comparison of the results from modelling to those of natural learners. Indeed, it had not even been established that humans would be capable of learning the task at all. Thus, in this study, healthy humans were tested on 4 variants of the 12-AX task, in each of which they had to learn the rules from trial and error and binary feedback within a single session of training.

The study aimed to test three main hypotheses: 1) Human subjects can learn the rules of a task like the 12-AX in a trial an error fashion in a single session. 2) Subjects can utilise additional structure provided by shaped sequential learning and trial and error learning benefits thereof. 3) Computational models of learning in the `12-AX` reflect learning in human subjects.

With respect to the first hypothesis, the study showed that healthy humans are readily capable of learning the `12-AX` task in a rapid manner. Over 80% of subjects learnt the full set of rules and subsequently were able to perform the task nearly flawlessly. They learnt the task on average in as little as ≈ 5 epochs, or less than 1000 stimulus presentations, corresponding to a wall clock time of about 15 - 30 minutes. This is much less than the model presented in previous chapters. However, the study also showed that the specific details of the experimental presentation played a large role in how well subjects learnt, as they performed substantially worse on the variant of the Apple-Axe, despite the rules being identical. Here, less than 30% of subjects learnt and

the average number of epochs needed for learning was more than double. Therefore, it was concluded, that the 12-AX task remains an appropriate and feasible task to use in future research into gated working memory. It lends itself well to experimental validation and provides the necessary breadth of difficulties to adjust to any given requirements. However, it also shows, current models of gated working memory do not yet capture the full flexibility of human performance.

The second main hypothesis postulated that a shaping protocol, as described in earlier chapters, helps humans to learn the rules of the 12-AX task more rapidly. Here, the study showed that shaping did have a positive effect and that overall subjects learnt better than subjects in the pure Apple-Axe task. The shaped version was derived from the more difficult Apple-Axe task rather than the easier `12-AX` version, to provide a more significant test of improvement for shaping as compared to the easier `12-AX` version. However, the results also showed that despite the additional information provided through shaping, the final task was still difficult to learn. Indeed, the higher fraction of subjects correctly learning the task compared to the unshaped Apple-Axe alone did not reach significance. Several of the non-learners, however, reached performance levels just short of the criterion and the overall learning curves were therefore significantly different.

It cannot be answered conclusively why the additional structural hints of the shaping tasks did not result in larger effects on the number of subjects reaching criterion. However, the debrief revealed that subjects could not fully recollect the individual stages of shaping and mostly did not identify the hierarchical relation between the shaping tasks and the full rule set. This hints that subjects might suffer from critical interference too, even though the task boundaries were clearly indicated and subjects were instructed to consider relations between the tasks. It also leaves room for exploring the benefits of other shaping protocols. Pilots were conducted on variations of the ordering of the shaping, and also tried switching tasks without telling subjects, but without results. Additional studies would be needed to asses whether shaping leads to different internal representations as predicted in chapter 4, and whether this leads to follow-on effects in e.g. generalisation or abstraction.

The third hypothesis, proposing that current computational modelling of learning in the `12-AX` task provides a full picture of human learning, can to some degree be refuted. At least on the behavioural measure of efficiency of rapidness of learning, current models do not reflect human performance. The exact definition of when an agent has learnt the task differs between the models and thus a direct comparison between models and with humans is difficult at best. Nevertheless, the order of magnitude difference indicates a significant discrapancy between the models and human performance. For example, results for learning in the LSTM showed the average performance at approx 160 epochs. Results for the PBWM model reported in O'Reilly and Frank (2005) were roughly 300 epochs. Todd et al. (2009)'s model of temporal difference learning in an gated working

memory actor critic model with eligibility traces reported a performance of nearly 2000 epochs under the best parameter settings and nearly 10000 epochs for lower eligibility trace settings. Even in the shaped condition presented in chapter 4, average learning times were at approx 40 epochs (median 14), although individual runs were significantly lower. In contrast, human subjects reached learning criterion after only 6 epochs for the `12-AX` case and still only 12 epochs in the Apple-Axe condition. Indeed, due to the limits of a single session experiment, the maximum possible number of subjects could learn for was 25 epochs. Thus, on the level of rapidness of learning, none of the models currently fully captures human behaviour.

In addition to this quantitative comparison between the learning rates of human subjects and computational models, qualitative comparisons can be drawn in order to begin to identify the mechanisms underlying different learning paradigms. The main contribution of PBWM and the models developed here, was the introduction of learning in a gated working memory model. The key prediction of these models states that a gated working memory exists and is used to solve the current task. In the case of PBWM, a further aim was to facilitate the anatomical mapping of these mechanisms. Since the current experiment was a purely behavioural experiment, the results presented here can not provide conclusive evidence for the existence of gated working memory, or for its mechanistic basis. Analysis of reaction times, however, was at least not contradictory to the concept of gated working memory, were it was shown that RTs are slower on trials requiring the remembrance of a stimuli than on those that are distractors and can be ignored. Importantly, however, by establishing the learnability of the `12-AX` task in a single session and determining appropriate parameters, the data pave the way for learning experiments in an fMRI scanner which may begin to address some of the mechanistic questions.

The categorisation of stimuli helps learning

The large differences in learning performance between the `12-AX` version and the Apple-Axe led to additional questions trying to identify the cause for this difference. One possible explanation was the extra hints available to subjects about the task structure through the strong categorisation of stimuli, i.e. that context markers were numbers and the sequences letters. This hypothesis was tested with the Spider-Train version. Although the data was not inconsistent with the idea that the categorisation helped learning, it could not explain the large difference between the `12-AX` and the Apple-Axe alone, and the improvement over the Apple-Axe did not reach significance.

The stimuli in the Spider-Train version were chosen amongst those available in this standard set to maximise this effect of categorisation, but it may still not have been as strong and salient as the difference between the numbers and letters. Indeed, dur-

ing the debriefing sessions, subjects did not comment on this categorisation on their own account in the Spider-Train version, as they did for the `12-AX` version. Additional experiments would be needed to either refute or confirm this hypothesis. For example, if categorisation matters, one might predict that a deliberate misalignment of the categorisation to the task structure may make the task even harder than with no categorisation structure in the stimuli.

Effects of working memory capacity on learning

An alternative hypothesis for the large difference in learning may be differences in working memory capacity, induced by the different stimuli (Alvarez and Cavanagh, 2004; Baddeley and Logie, 1999; Cowan, 2001). Any such differences in WM capacity of the abstract pictorial representations of simple objects, compared to the letter and number stimuli of the `12-AX` version, might contribute to variations in learning difficulties. These were not big enough though, to influence steady state performance after learning. Neither the overall error rate, the error rate increase with distance to the context markers, nor the reaction times were significantly different in the assessment phase between the `12-AX` and the other three versions. However, the ideas of selectively gated working memory are precisely to minimise the need for excessive working memory. Therefore, before such gating has been learnt, the demand on WM capacity may well be more important, amplifying any effect of capacity.

Although no direct evidence for this phenomenon exists, observations in various modelling work may support this notion. For example, Todd et al. (2009) showed that the eligibility trace parameter λ, a parameter influencing the decay of the state trace memory, heavily influenced the time his model took to learn the 12-AX task to criterion. Similarly, in chapter 4.5 the frequency of context markers was varied, which may be said to vary the necessary demand on memory capacity during learning. This variation also had a strong effect on training times. In future studies it will be important to identify and investigate potential correlations between individuals' working memory capacity, the WM load induced by the stimuli and the per-subject training time. Indeed, with respect to the hypothesis in chapter 4.5, it would be interesting to test the prediction that subjects who went through shaping would be less sensitive to memory demand and capacity during learning. This would be expected since the learnt gating of relevant stimuli during early stages of shaping would reduce the need to memorise intermittent stimuli.

Motivation and the limits of a single session study

One of the conclusions drawn from this study is that the Apple-Axe version is more difficult to learn than the 12-AX version, and that only a relatively low percentage

of subjects achieved the criterion and could afterwards name the rules. It is however important to remember, that this conclusion can only be drawn within the parameters of the experimental setup, particular the constraint of learning in a single session. This includes the limit of training for no more than 25 epochs and taking no longer than two hours overall for the experiment. Whether more subjects would have learnt the task if they were given more time, remains an open question. Indeed, this limitation means, that almost all runs of any of the models described in the other chapters would have had insufficient time to reach the learning criterion. This obviously also limits the distribution of epochs taken to learn the task to 25, potentially introducing a form of sampling bias for the speed of learning, when compared to the much longer capped (500 epochs) results of chapter 4. Whereas for the `12-AX` version, the distribution stayed well clear of this limit, and therefore appeared to not effect it much, the same can not be said for the Apple-Axe or Spider-Train version. It cannot be ruled out, that average learning times for the Apple-Axe would be as slow in an unconstrained setup as those seen in models, although this appears to be unlikely. The two hour limitation is perhaps even more severe for the shaped Apple-Axe version, as it additionally includes the 3 pre-tasks, leaving less time for the final, full, task. Whereas in the three other conditions, the two hours were mostly sufficient time to reach the 25 epochs (average number of epochs trained for subjects not reaching the criterion was 23.5, 24.2, 23,4 for the `12-AX`, the Apple-Axe and Spider-Train version respectively), the number of epochs spent by subjects who failed to learn the task in the Apple-Axe stage was only 19. Three subjects even spent less than 17 epochs in the final full stage. This factor might have contributed to the relatively low success rate, but simultaneously higher performance of non-learners. The two hour session maximum, however, already proved to be rather long, with several subjects showing signs of lack of concentration and frustration by the end of this period. Of cause, the split into several tasks may have helped in reducing the monotony.

Despite these motivational issues potentially being rather important (e.g. Kouneiher et al., 2009; Savine and Braver, 2010), they nevertheless had to be disregard as an additional sources of noise. Apart from the measures taken (time limit and performance related pay) the study could not control or measure motivation explicitly. Equally, the models presented here do not have the ability to capture such effects either. To limit its influence, any longer training would have required a multi-session training set-up. Indeed, for example Todd et al. (personal communication 2009) chose to use a 10 session setup for learning an artificial grammar-like task not dissimilar to the 12-AX. They observed under some conditions gradual learning over the full duration, although at the time of communication it was not yet conclusive when such gradual learning occurred. The choice of single session training, however, also allowed to avoid confounds such as inter-day, reconsolidation and insight effects (Wagner et al., 2004; Diekelmann and Born, 2010) or simply uncontrolled communication. It was therefore more appropriate for an initial study on overall feasibility of learning the 12-AX task.

The single session training paradigm is also likely to be better suited for potential future follow-up studies, such as fMRI or TMS studies and so it is important to have a comparable behavioural basis.

Anatomical considerations of learning in the 12-AX task

As a purely behavioural experiment, this experiment could not attempt to address the interesting aspect of the localisation of brain areas / networks involved in learning the 12-AX task. Models like Frank et al. (2001b) Prefrontal cortex, Basal ganglia Working Memory model (PBWM), and to a lesser degree (and more implicitly), the models presented in chapter 4, make specific predictions of a mapping to brain structures. There is ample evidence in the literature (some of which is reviewed in chapter 2) suggesting that parts of the PFC as well as the basal ganglia would likely be involved in learning the 12-AX. Indeed, the fMRI experiments on instructed subjects of Reynolds (2007) showed that dlPFC was differentially active to sequences based on the context in the 12-AX. Likewise, similar tasks have been used to probe differential activation in PFC (Koechlin et al., 2003; Koechlin and Summerfield, 2007). However, it would be interesting to get direct verification for this during learning and test if the potential representational changes of shaping can be identified with methods such as fMRI or TMS.

Associative learning vs hypothesis driven learning

As hinted at in Eshel et al. (2009) with the AX-CPT, subjects might employ different strategies to learn and perform 12-AX. Without memory subjects should not perform above chance. However, when they applied TMS over dlPFC, subjects slowed down but were not impaired in performance. One possible explanation Eshel et al. offered, was that different strategies involving different brain areas, such as the use of episodic memory rather than working memory, may substitute for the impaired system. This is perhaps similar to the well studied trade-offs between hippocampus and basal ganglia in place vs response learning (e.g. Packard et al., 1989; Packard and McGaughy, 1996; Devan and White, 1999), or the dual controller hypothesis described by Daw et al. (2005), involving goal-directed vs habitual control. The latter of which is thought to be located in PFC and the basal ganglia respectively (Anderson, 1982; Dickinson, 1985; Sloman, 1996).

Different strategies like goal-directed vs habitual learning may also underlie some of the considerable differences in efficiency of learning 12-AX between humans and the sample computer models. The LSTM models presented in chapter 4 typically required between 100-200 epochs. However, humans learnt the `12-AX` task on the order of 10 epochs. Even the hard versions like the Apple-Axe could still be learnt in less

than 25 epochs. It is possible that humans embark on a form of hypothesis testing, i.e. try and learn the rules by postulating a set of rules and testing to see if they are consistent with the task responses. This strategy is clearly not compatible with how any of the current theoretic models of learning the 12-AX work, which rely in one form or another on slowly strengthening statistical associations. Indeed, these two strategies may both be active simultaneously in humans with a preference for hypothesis testing where successful. Unlike experiments such as Gläscher et al. (2010), which used a combination of behavioural modelling of both, goal directed and habitual as well as fMRI correlational analysis, to determine when subjects use one or the other strategy, the present study provided no such opportunity, for the lack of a good model of hypothesis testing in the 12-AX task.

One potential other, more indirect, way of assessing such strategies might be timing. As suggested in Daw et al. (2005), one of the reasons for not always using the goal directed controller is computational complexity of forward planning in the goal-directed search. Similarly, hypothesis testing may require more time and one could hypothesise that subjects would be slower with such a strategy. By forcing subjects to respond quickly rather than allowing for a self-timed paradigm as this study did, the hypothesis would suggest that subjects may switch strategy. However, even unforced, subjects responded quite quickly and showed reaction times of only about 1 second on average during learning, making it hard to rely upon a forced timed paradigm. Nevertheless, a considerable trial-to-trial variability existed, with a few long response times. In an fMRI experiment it would be interesting to see if long reaction times would correlate stronger with stimulus-independent thought for planning and hypothesis generation, while short RTs correlate with stimulus-oriented thought (Burgess et al., 2007a).

Furthermore, the selection of instructions given to subjects may have influenced the outcome of learning (Doll et al., 2009) and the strategy chosen. In the current experiments, subjects were given very elaborate sets of instructions (A.2) with detailed properties of the rules to learn. Instructions were deliberately chosen to include as much detail as they did in order to limit the set of rules tried by the subjects to those that the models were equally capable of. In addition it was important to overcome motivational issues in order to not ignore low but non zero error rates. However, the strong emphasis on specific properties of rules may have instead inadvertently shifted the subjects' learning strategy towards an explicit hypothesis-driven approach, thus making results less comparable to the models. In addition, the perceived complexity of instructions may have had an effect on the hypothesis entertained by subjects. For example Goodman et al. (2008) postulated that humans have a prior on hypotheses with a systematic complexity bias, following similar ideas of Feldman (2006). Should this bias be affected by the instructions, the speed of learning could also be affected. Indeed, one subject who had learnt the rules reported during the debrief that the rules turned out to be simpler than they had expected, and that they were thus not sure if they had missed something more complex. Without a systematic repetition of the ex-

periments with different instruction sets, however, it is not possible to determine their influence and particularly the instructions interaction with shaping. However, given the self-timed nature of the experiment and thus the likely usage of some explicit strategy, detailed instructions were deemed to be overall beneficial in reducing variance.

Despite the complexity of the instructions, subjects had no problems comprehending the instructions as was verbally verified during the familiarisation phase. This was potentially due to the fact that all subjects had undergone many behavioural experiments previously, providing them with an inherent familiarity with the necessary concepts. Whether this verbal comprehension of the instructions meant, that the subjects had fully internalised the rules during the experiment, however, could not be determined. It is thus not impossible that subjects were affected by goal-neglect (Duncan et al., 1996), an effect in which subjects can exactly say what they should do, but then during the task fail to act upon it. Due to severe capacity limitations of humans to compile verbal rule instructions into executable task sets (Dumontheil et al., 2011), goal-neglect increases with rule complexity (Duncan et al., 2008). Interestingly, it was the overall complexity of instructions that was relevant in these experiments, rather than the complexity of the rules that were currently active. As apparent model complexity can be reduced by familiarity and training shared across multiple tasks of similar nature, the frequent exposure of subjects to behavioural experiments may again reduce this effect.

Overall, these results confirm that the 12-AX task is suitable for behavioural testing. It is a good choice for future empirical linking of predictions made by the various models of sequential learning. Task complexity is well within the learnable range of humans, yet sufficiently high to reveal interesting structure. This supports the continued use of the task, both from a theoretical as well as experimental view, although the results so far have also shown some striking differences between the two, particularly the efficiency of learning. It is thus necessary to keep these differences in mind when comparing to human performance and to reconcile them before making direct behavioural predictions.

Chapter 6

Separation of rules and stimuli, a need for variables

6.1 Introduction

Chapter 4 presented several arguments for how the sequential learning of compound structure can help cognitive flexibility in the form of more rapid learning of new tasks. Further, it showed that such structured learning can also result in improved internal representations, helping the model perform better in various secondary markers of cognitive flexibility, such as abstraction and generalisation. However, one important aspect not covered was the treatment of stimulus rule abstraction. Chapter 4.8 did look at reversal tasks with repeated switches back and forth between the same two variants of the task. But it did not look at the more challenging version of repeated switches to as yet unseen versions of the task; i.e to keep the same rule set, but to use a different mapping to the underlying alphabet of percepts. Even then it was evident that switches within the same stimulus alphabet would be problematic, questioning the ability to abstract rules away from their physical incarnation in the world. Nevertheless, from every day life examples, it is clear that such abstraction represents an important element of natural cognitive flexibility.

There are at least two interesting and relevant sub-aspects to the question of how such rule abstraction might be handled: On the one hand is the question of learning, e.g. does abstraction occur immediately on initial learning, or does it require a gradual re-representation to achieve abstraction over repeated exposure to switches? On the other hand is the question of steady state execution and switching to tasks for which the agent already knows the rules. That is, how should tasks and rules be represented in a neural architecture, that it can rapidly switch with changed stimuli?

These questions of rule to stimulus abstraction can be seen as linked to the wider idea of "conceptual knowledge" that has been studied in various forms (e.g. Murphy, 2004).

Although the term "concept" is used in many different ways, one particularly important one is in terms of "the capacity to respond in the same way to a variety of different stimuli". In recent years the interest of "conceptual knowledge" has increasingly focused on the capacity to generalise and facilitate learning in general decision making (Shea et al., 2008). However, so far little is known about the neural systems and mechanisms underlying such abstracted decision making.

Kumaran et al. (2009) nicely demonstrated the human ability to abstract concrete representational stimuli from higher level task structure. In a variant of the weather prediction task (Knowlton et al., 1994), which used structured shape-shape or shape-location conjunctions rather than simple elemental information, Kumaran et al. showed that humans learn to predict faster in a second session, if the structure remained constant, but the stimuli switched to new so far unseen ones, compared to initial learning with unknown structure and stimuli.

The Wisconsin Card Sorting Task (WCST) focuses even more directly on the question of the switch of stimulus, while keeping the structure of the task unchanged. In the WCST, the rule to sort cards remains constant over time, however, the dimension according to which the rules apply switches throughout the task. In this setting, healthy subjects have no difficulty immediately adjusting to the new dimension without having to re-learn the rules. Even pre-frontal damaged patients, who typically perform much worse at this task, tend to show perseverative errors rather than random errors (Sullivan et al., 1993; Milner, 1963). This suggests that their rule knowledge and execution remains intact, and is not unlearnt and relearnt during switching. Even though typically the WCST is used with a very limited set of stimulus dimensions, and can be seen to resemble more the switching paradigm presented in chapter 4.8, there is little evidence to believe subjects would perform worse if switches continuously involved new dimensions.

In the past some authors have argued that at least the full idea of conceptual knowledge requires the inclusion of language abilities (Chater and Heyes, 1994; Bermudez, 2006), and is thus limited to humans. Due to a lack of language capabilities, this would predict that the type of models considered here would be unsuitable. However, the separation of rules and stimuli can be found to some degree in animals, too.

A particularly striking example is the study by Shima et al. (2007), in which monkeys were trained to perform various motor sequences. Each sequence of four movements followed a pattern of either AAAA, AABB or ABABA, with three different possible actual movements. Even though it would have been possible to learn each sequence independently, Shima et al. found that a large number of prefrontal neurons showed strong modulation to the categorical or abstract structure instead of the individual movements or sequences. Thus, these neurons are able to separate sequence rules from concrete movement execution.

Equally, the delayed (non) match to sample task, commonly used to probe working memory with monkeys (e.g. Miller et al., 1991, 1996), incorporates this separation of rules and stimuli. Whereas the rules remain constant, i.e. to respond to a stimulus in a sequence that matches the cue stimulus, the actual stimulus changes each time, needing to bind the variable stimulus to the rule. Here too, single neurons in the PFC encoded for the selectivity of the abstract rule, independent of cue (Wallis et al., 2001). While one common scenario uses only a few highly familiar stimuli, allowing monkeys to simply learn each pattern individually, experiments with large numbers of novel stimuli (Miller et al., 1996) have confirmed monkeys' ability to also represent the task conceptually.

Such abstraction, however, does not emerge easily from typical connectionist models. For there, the task rules are encoded in the weights of individual neuronal elements such that the encoding and internal representation of stimuli are inherently linked with the rule structure. If the encoding of the stimuli changes, typically the rule representations in the network are entirely lost as well.

One common remedy to aspects of this problem can be a form of generalisation or "reduced descriptions" (Hinton, 1990). By having part of the encoding shared between stimuli amongst which a switch should be possible, encoding of the rules can generalise to this shared encoding. Indeed, categories might often be defined to combine those elements with shared properties to facilitate abstraction.

However, it is clear that natural learners can also realise rules and tasks built on abstract categories, such as "a context marker" to which arbitrary objects or stimuli are assigned on an ad hoc basis. In this setting of arbitrary and repeated remapping of stimuli, which is the focus of this chapter, a static representation is not sufficient. Instead, a form of *variable substitution or mapping* is likely to be necessary to achieve the full flexibility, such as in the case of the ABAB rule mapping onto the push-pull-push-pull movement in the experiments of Shema et al.

This chapter presents an analysis and an example model of how rules and stimuli can be separated, allowing for rapid switching. Building upon the framework of previous chapters, it shows that the principles of gated working memory models can readily be extended to realise an abstract rule representation.

As variable substitution is likely to be fundamental in many forms of advanced cognitive processing, a variety of approaches has previously been treated. These included for example BoltzCONS (Touretzky, 1990), featuring a sophisticate network that can implement symbolic linked lists and tree structures, and a simpler model by Moody et al. (1998), solving the delayed match to sample task with a recurrent network. Key to its operation are a set of storage units, as well as a specific comparator network. Both components are also integral in the model presented.

It has been suggested that a hierarchy exists in PFC, along which ever more abstract

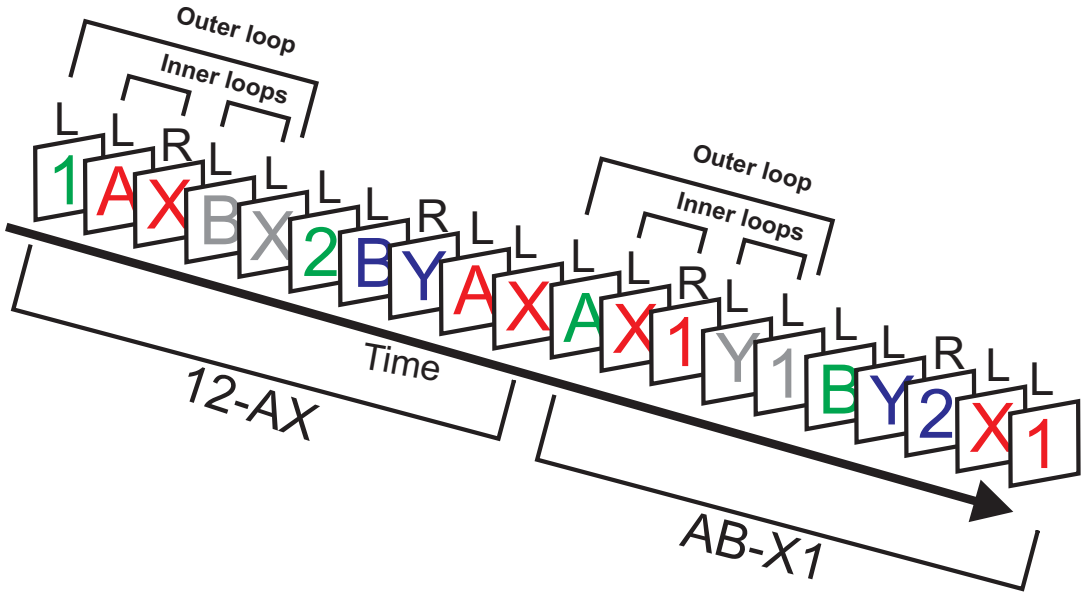
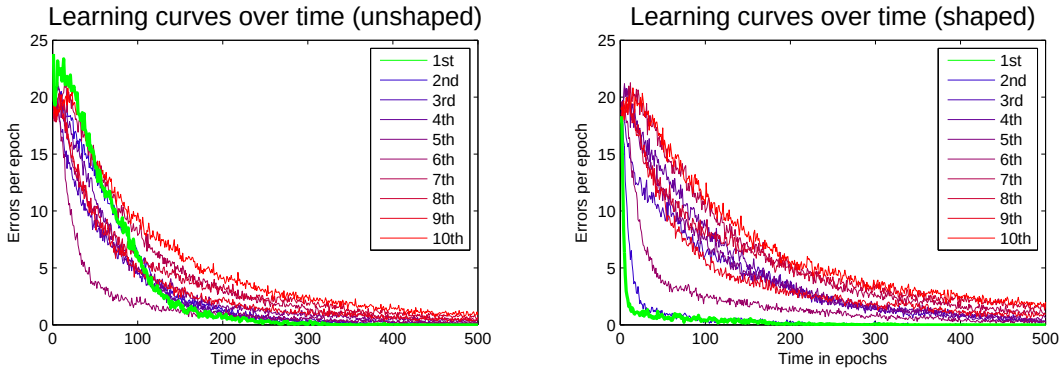


Figure 6.1: Generalised 12-AX: The generalised 12-AX task uses the same rule structure as the standard 12-AX task with its hierarchical context markers and inner loop target sequences. Occasionally the stimulus to rule mappings change. This example shows a switch from the standard `12-AX` to the `AB-X1` task, where `A` and `B` are the context markers and `X1` and `Y2` are the target sequences in the respective context.

rules are represented. While neurons in PMd show elevated firing during response relevant cues, dorsolateral PFC show a model group of neurons firing independent of cue put modulated by the abstract rule currently active. In contrast to previous accounts, an explicitly layered approach to abstraction was therefore taken, creating an additional level of indirection. Nevertheless, the core computational functionality remains within the class of gated working memory, allowing to scale previous models to easily include rule - stimulus abstraction.

Variable substitution was studied in a generalised version of the 12-AX task. As shown in figure 6.1, here, once an agent has learnt the `12-AX` task, the task switches to a new task, such as the `AB-X1` task. Its abstract rules remain identical, i.e. contain two context markers that switch between two potential target sequences of two stimuli each. However, the stimuli used are different. So in the `AB-X1` case, it would be `A` and `B` for context markers and the target sequences `X1` and `Y2`.

As an extension to the analysis of reversals in chapter 4.8, it is first shown that the standard LSTM model does not trivially exhibit the behaviour of separation, and does not show more rapid adaptation to repeated switching. Next, the abstract bi-linear model demonstrated that an extended internal representation allows a gated working memory model to show such rapid switching once learnt. Finally, it is shown that this idea can be transferred to the LSTM model, although no model of learning is presented here in this initial work.



(a) (b)

Figure 6.2: LSTM performance on the generalised 12-AX: Averaged learning curves for each of the 10 rule-stimulus mapping switches aligned to the beginning of a rule-stimulus mapping switch. Data shows that there is no correlation between rapidness of learning and how many switches have occurred. (a) shows data for unshaped networks. (b) shows data for the shaped networks. The first two rule mappings were shaped beforehand.

6.2 Stimulus - rule abstraction through a memory layer of indirection

Standard connectionist models often fail to show stimulus-rule abstraction. For example, the LSTM network presented in chapter 4 did not show any sign of improvement over repeated switches in the generalised 12-AX, as shown for example in figure 6.2. As expected, shaping alone did not help to remedy this shortfall either. A plausible model to demonstrate some of the components and aspects that may underlie such flexible behaviour was therefore tested. In order to focus on the question of how an architecture may be structured to enable this type of behaviour, the model was restricted to *steady state execution*, postponing the question of learning. Furthermore, a more abstract architecture than the LSTM was chosen, to concentrate on the mathematical properties rather than the immediate biological plausibility, especially as the LSTM lacks a key feature to implement such a schema efficiently, as discussed in section 6.3.

6.2.1 A suggested model

The model presented here heavily draws upon the rule based part of the bi-linear model introduced by Dayan (2007) and reviewed in chapter 3.4. Similar to all gated working memory models, it features N dedicated working memory components, whose contents are each controlled by a gating mechanism. In this case, the gating is provided by binary decision units driven by a probabilistic soft-max function on a bi-linear weighting of the input and working memory state space.

The structure of the model is shown in figure 6.3, which consists of a set of rules that

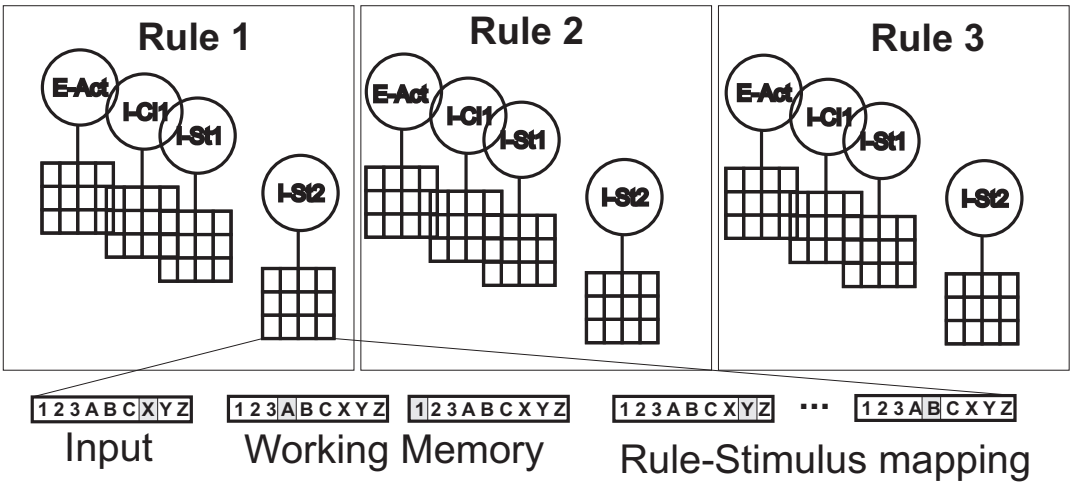


Figure 6.3: The bi-linear model: This figure shows the overall decomposition of the extended bi-linear model. The full task is broken down into several individually simple rules. Each rule consists of a set of binary decision units for both internal and external actions whose activation are calculated by a standard bi-linear form from the state vector. Internal actions are "store" and "clear" for each of the memory units. The state vector including the current stimulus and the content of working memory is extended by additional memory units storing the rule stimulus mapping.

jointly encode the full task. Each such rule controls a set of (binary) units which drive both, external actions and the internal memory gating. As at each time step all rules are executed simultaneously, an additional overall output gate for each rule exists, that arbitrates between the rules and determines which rule drives the actual behaviour.

Execution of each decision unit is according to a standard bi-linear function:

$$P(o^c = 1) = \sigma\left(\sum_{ij} x_i W_{ij}^c x_j + \sum_i u_i^c x_i + b^c\right) \tag{6.1}$$

Although each rule could in principle feature the full complexity of the bi-linear mapping, the task is broken down into a set of simple logical rules, featuring simple predicates like "*and*", "*or*" and "*equals*". As shown in table 6.1, the 12-AX task is broken into 10 separate rules, each in a simple conjunctive form that results in a linearly separable logic function.

The model described so far equally encodes the rules inseparably from the representation of the stimuli and therefore suffers from the issue of inflexible rule - stimulus mapping. To overcome this limitation, the introduction of a layer of indirection into the model was proposed, allowing to separate the encoding of the rules from the encoding of the symbols, and thus to change one without the other.

To this extent, additional working memory modules are included in the model that hold the mapping of rules to stimuli. For each abstract stimulus function an addi-

Rules:	External Actor	Internal S-1	S-2	C-1	C-2
$input = 1$	0	1	0	1	1
$input = 2$	0	1	0	1	1
$input = A$	0	0	1	0	1
$input = B$	0	0	1	0	1
$input = C$	0	0	1	0	1
$input = X \wedge mem - 1 = 1 \wedge mem - 2 = A$	1	0	0	0	1
$input = Y \wedge mem - 1 = 2 \wedge mem - 2 = B$	1	0	0	0	1
$input = X \wedge \overline{(mem - 1 = 1 \wedge mem - 2 = A)}$	0	0	0	0	1
$input = Y \wedge \overline{(mem - 1 = 2 \wedge mem - 2 = B)}$	0	0	0	0	1
$input = Z$	0	0	0	0	1

Table 6.1: Training rules: This table shows the list of rules into which the 12-AX task is broken down. If the condition of the rule is met, the 5 decision units of the rule fire according to the table. The external actor determines the L/R response. S-1 and S-2 represent the store units for the two working memory modules. C-1 and C-2 represent the clear units. The sixth unit, the enable unit, fires according to if the rule is triggered.

tional memory module is needed. In the case of the 12-AX task, this would be 9 additional modules. Two for context markers, four for the target sequences and the rest as distractors. With each memory module loaded with the respective representation of stimulus mapped to its abstract function, rules can be adjusted to include the necessary abstraction.

With this structure in place, rules can be encoded in the weights of the network, independent of the actual encoding of stimuli. Instead of simple rules of the form "$input = 1$" or $input = X \wedge MemoryCell^1 = 1 \wedge MemoryCell^2 = A$, rules now become "$input = MemoryMapping^{(1)}$" or $input = MemoryMapping^{(X)} \wedge MemoryCell^1 = MemoryMapping^{(1)} \wedge MemoryCell^2 = MemoryMapping^{(A)}$". In this form, rules no longer depend in any way on the actual encoding of stimuli, but solely on the equality of contents of its memory modules. Thus, switching between e.g. the `12-AX` and the `AX-C2` becomes a matter of updating the respective memory modules and can be achieved in a single step, assuming the new mapping is known. Given that the bi-linear function is ideally suited to determine equivalence or equality between its memory cells, the model is easily able to encode the adapted set of rules.

6.2.2 Training the model

The overall training regime occurred in accordance with Dayan's original model. Each set of weights was trained in a supervised, non-sequential manner, using standard gradient decent.

For these simulations, a network configuration of two working memory modules was chosen. In addition to cope with the rule - stimulus mapping, an extra 9 memory

modules were added. They were, however, not gated by the network but remained fixed with the respective values for the mapping. As such, each rule consisted of 6 binary units: An execution unit, an external actor and a store and clear unit for the two gated working memory modules. The state vector $\vec{x}$ was a conjunction of the input stimulus, the two gated working memory units and the 9 additional rule-stimulus mappings. The stimulus encoding was chosen to be a simple unary line labelled code with values 0 and 1. Thus, the state vector overall was a 12*9=108 dimensional vector. This encoding was chosen to be consistent with the LSTM network, but other encodings, such as a binary encoding, could equally be used. During tests, the encoding did not appear to have any significant effects, other than computational complexity and a reduction of the number of weights in the network.

Two different groups of training examples were generated:

In the first group, whose results are described in section 6.2.3, training examples were generated with random rule - stimulus mappings. Each training example state vector was generated by first randomly sampling a permutation of the stimulus alphabet as the rule - stimulus mapping part of the state vector, followed by a randomly sampling a stimulus for each, the input and the two memory cells. If a uniform distribution were chosen for the sampling of memory cells, the vast majority ($\approx 90\%$) of state vector configurations would never occur, if the task were performed correctly and the correct information gated into memory. To reduce this inefficient use of training examples, an over-weighting of valid configurations (i.e. those configurations, whose memory content would occur in a correctly gated model for the given rule mapping) was chosen, with a ratio of 9:1. The supervised training outputs were generated according to each of the rules. Network weights were trained with a set of 5000 training examples.

The second group, presented in section 6.2.4, used a fixed rule-stimulus mapping, that always induced the `12-AX` task. Other parameters of training example generation were identical, although due to the limited number of distinct training examples of 729 a smaller set of training examples was used to exclude duplicates.

Training of the network weights were performed, using a quadratic loss function on the probability of the soft-max function. Minimisation was performed with the Matlab function minimise [1], which uses a Polack-Ribiere flavour of conjugate gradient decent. The temperature parameter of the soft-max was chosen higher ($\beta = 1$) during learning than during execution as a form of annealing schedule.

Overall, this model had a large number of parameters. With 10 rules and 6 units per rule, there were overall 60 bi-linear units. Given a state space vector dimensionality of 108, each bi-linear unit contained 11773 weights, resulting in a total number of 706380 weights. Although some options exist to reduce the number of parameters, such as sharing weight matrices between units encoding the same function, they only provide

[1] http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/

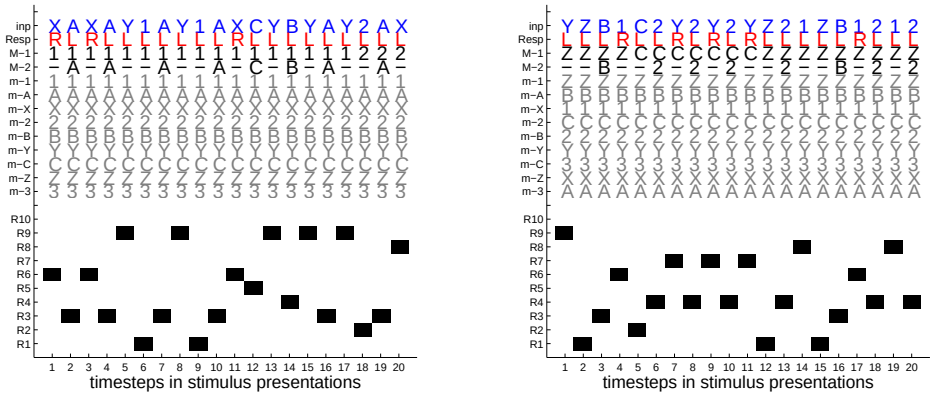


Figure 6.4: Execution in the bi-linear model: This figure depicts a sample run from the bi-linear model, showing the content of the mapping table as well as which rule fires for each stimulus. M-1 and M-2 represent the contents of the two WM modules. Inp represents the input and Resp the output response. R-1 to R-10 show which rules fire to any given time. (a) shows the execution of the 12-AX task. (b) shows the execution of the ZC-B1 task. Despite different stimuli, the same rules keep firing as in the 12-AX.

a reduction by a constant factor. Nevertheless, with the simplicity of the individual rules, most weight matrices were highly sparse and manageable, even though the ratio of weights to training examples was high. To foster sparseness and potentially help generalisation performance, an l_1 regulariser was included during training.

6.2.3 An effective solution for execution in the generalised 12-AX

After training the network with a full set of training examples, the model successfully learnt the generalised 12-AX task, correctly switching between arbitrary rule - stimulus mappings. Conditioned on a correct mapping loaded into its working memory models, the model performed error-free over the course of testing it on the 10 random permutations in the generalised 12-AX task, for 30 epochs each. As shown in an example run in figure 6.4, the model correctly fired each of the appropriate rules and adjusted its behaviour once a different set of mappings (e.g. the ZC-B1 task) was loaded (figure 6.4b).

Analysing the resulting weight matrices (sample weights for a set of rules are shown in figure 6.5) shows a distinct pattern emerging. While rule weights for the non-generalised 12-AX (figure 6.7, Dayan (2007)) resulted in only a few individual non-zero weights, the learnt solution to the generalised 12-AX had a distinct structure. As the state vector itself was a concatenation of encoded input and memory vectors, the weight matrix subdivided into a set of blocks, each block providing a multiplicative operation between two parts of the state vector. For the generalised 12-AX, all sub-blocks came close to diagonal or scaled identity matrices, akin to the structure shown in figure 6.6.

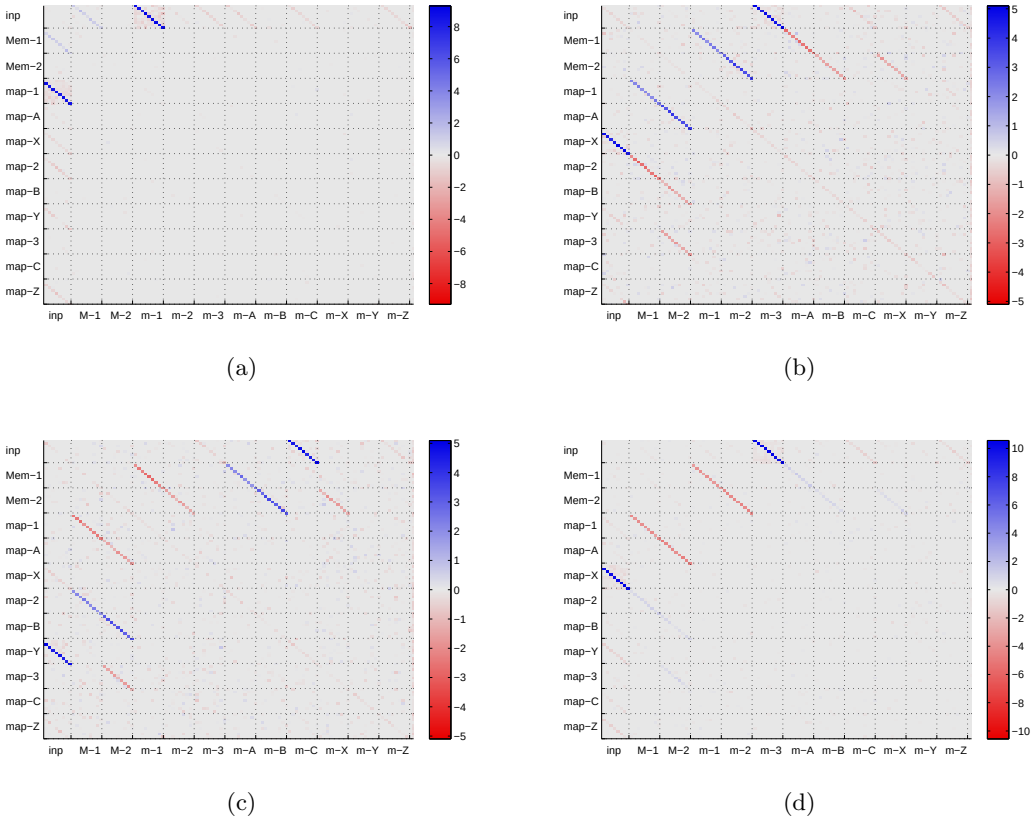


Figure 6.5: Example weights trained on the generalised 12-AX task: These graphs depict the weight matrices for the enable units of four different rules, drawn from a typical run. Weights were trained in a supervised way from training examples generated from the full generalised 12-AX. The rules are $input = 1$, $input = X \wedge mem-1 = 1 \wedge mem-2 = A$, $input = Y \wedge mem-1 = 2 \wedge mem-2 = B$ and $input = X \wedge \overline{(mem-1 = 1 \wedge mem-2 = A)}$ respectively.

It is termed this multi-diagonal.

The structure of a multi-diagonal matrix sub-serves the important operation of matching or comparison. Under the assumption that the stimuli are encoded with a strategy limited to 0/1 or -1/1 per dimension (here 0 / 1), only if the two parts of the state vector (s_1, s_2) for a given block b are identical, is the sum of the block $(\sum_{i=\vec{s_1},j=\vec{s_2}} W^b_{ij} x_i x_j)$ greater than 0. Therefore, the sign of each block provides an equality test, resulting in an overall linear function of equality operations. Basically any other form than the multi-diagonal matrix would not be agnostic to the underlying encoding of the stimuli, and thus would not be able to provide the necessary abstraction to achieve instant remapping without the need to alter weights.

Rule based vs habitual behaviour in the bi-linear model

An important point in Dayan (2007) was the interaction between the rule based learning and habitual performance, which were both sub-served by the same computational

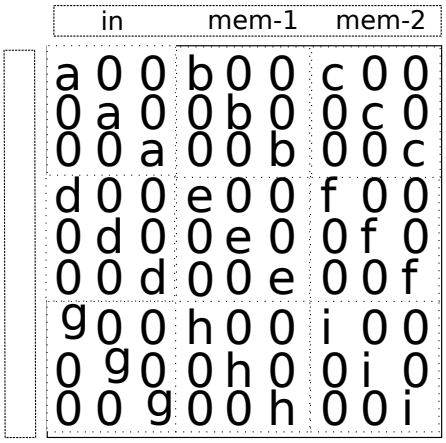


Figure 6.6: Multi-diagonal restriction: This cartoon shows the structure of the restricted multi-diagonal matrix. Due to its structure it allows to test for equality between two elements of the state vector.

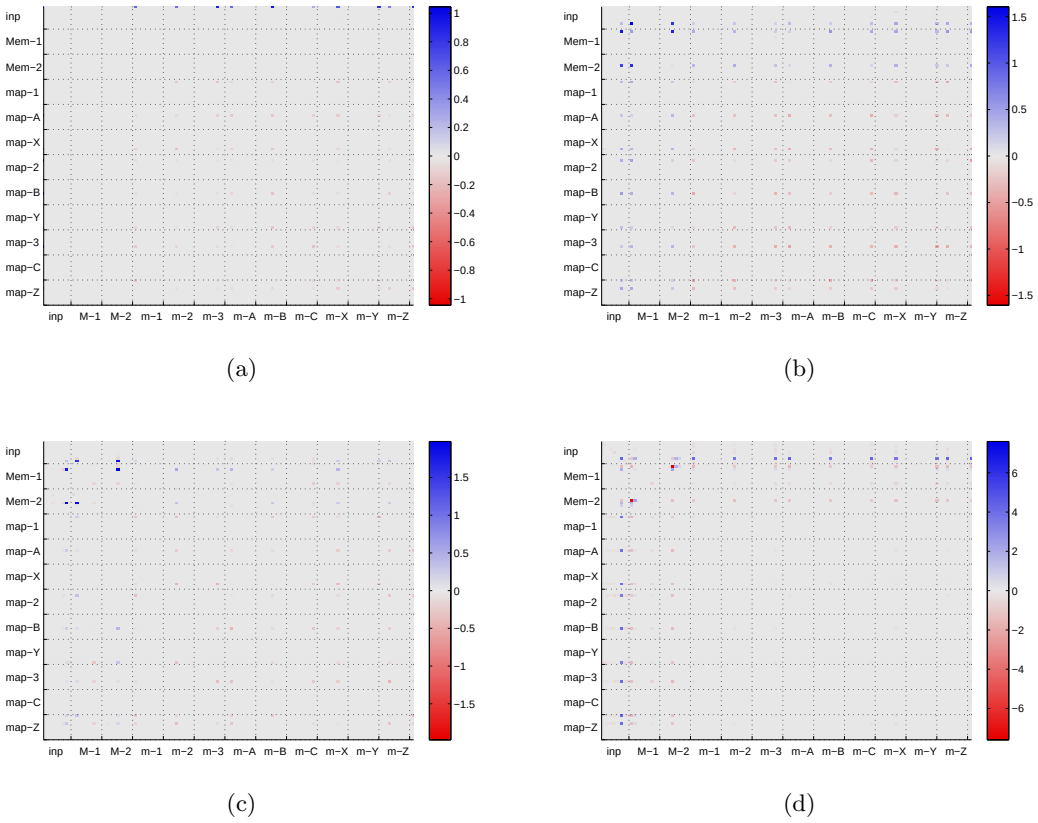
substrate. However, instead of multiple simpler rules, each potentially driving behaviour, habits were encoded in a single, albeit more complex bi-linear form. So far, this account purely focused on the rules' point of view. To analyse whether the rule - stimulus abstraction extends into the regime of habits (as defined by Dayan), a single weight matrix was trained on the full generalised 12-AX.

None of the runs were successful in reducing training error to 0, showing that in its current form stimulus rule abstraction does not carry over to the habitual case. Analysing its behaviour, it is clear that the bi-linearity is computationally not sufficient to encode the full rule set of the 12-AX task, while still providing the stimulus abstraction through matching inputs to contents of working memory. As described above, by its nature of being agnostic to the input encoding, the applicable weight structure reduces to a multi-diagonal form. As such, its computational complexity translates into the equivalence of a linear model on the transformed space of (non-)matching memory mappings. The 12-AX task, however, is not a member of the class of linearly separable functions.

6.2.4 Automatic generalisation

For the results in the previous section, all network weights were trained on a large variety of different mappings, forcing the network to operate through the layer of indirection and thus achieving the desired flexibility and abstraction. One important question, though, is whether such flexible representations and rule encodings emerge immediately with the first switch, or whether extensive exposure to a variety of different mappings is required. From human experiments like the ones by Kumaran et al. (2009), it is clear that to some degree flexibility and abstraction emerge immediately, although their effect might still strengthen over repeated exposures.

To see if the same is true or achievable in the context of the current model, a set of



(a) (b)

(c) (d)

Figure 6.7: Example weights trained on the non-generalised 12-AX task: These graphs depict the weight matrices for the enable units of four different rules, drawn from a typical run. Weights were trained in a supervised way from training examples generated from the full non-generalised 12-AX. The rules are $input = 1$, $input = X \wedge mem-1 = 1 \wedge mem-2 = A$, $input = Y \wedge mem-1 = 2 \wedge mem-2 = B$ and $input = X \wedge \overline{(mem-1 = 1 \wedge mem-2 = A)}$ respectively.

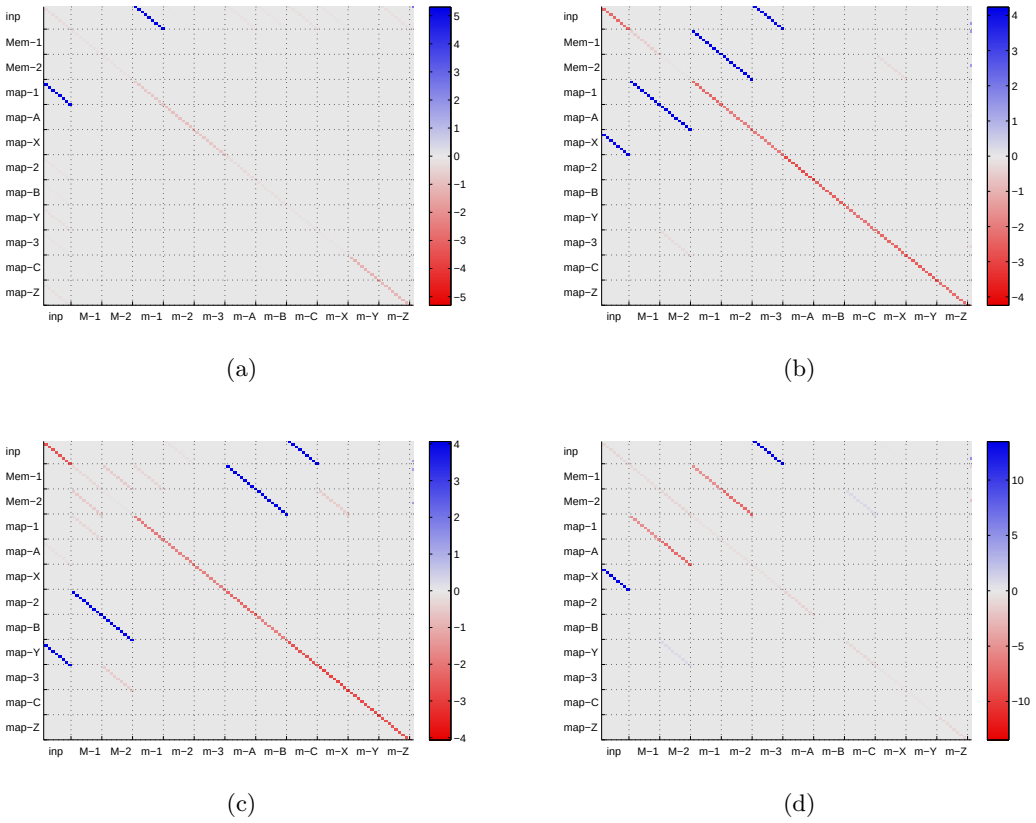


Figure 6.8: Example weights for the multi-diagonal restriction trained on the non-generalised 12-AX task. These graphs depict the weight matrices for the enable units of four different rules, drawn from a typical run. Weights were trained in a supervised way from training examples generated from the full non-generalised 12-AX, but limited to the multi-diagonal form. The rules are $input = 1$, $input = X \wedge mem - 1 = 1 \wedge mem - 2 = A$, $input = Y \wedge mem - 1 = 2 \wedge mem - 2 = B$ and $input = X \wedge (mem - 1 = 1 \wedge mem - 2 = A)$ respectively.

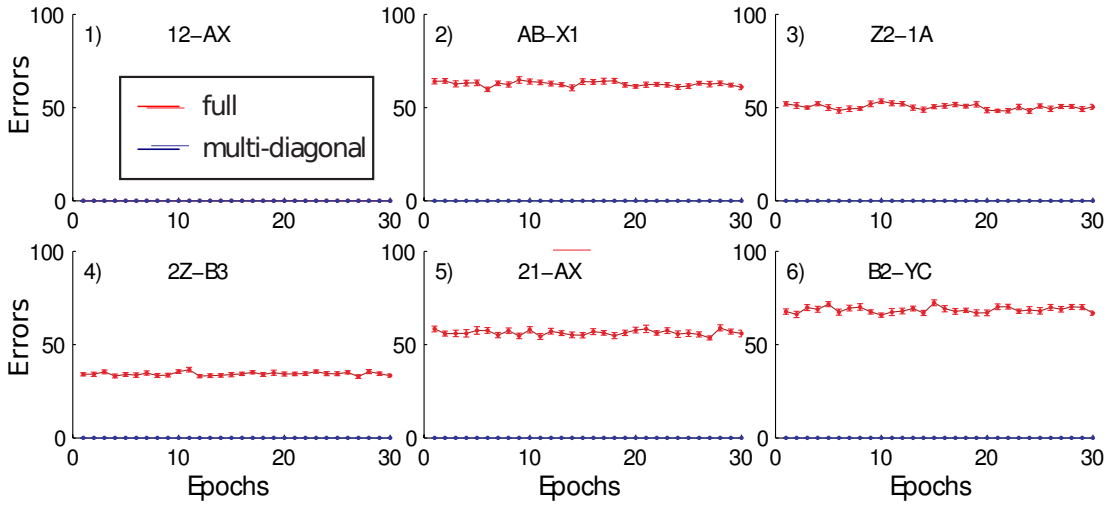


Figure 6.9: Automatic generalisation: This graph shows the error rate of executing the generalised 12-AX task, comparing the full rank matrix model to the multi-diagonally restricted matrix. Both are trained on samples of the 12-AX task alone. Both models show perfect performance on the trained 12-AX task. However, unlike the full matrix model, the multi-diagonal model also shows perfect switching behaviour despite being trained only on examples of the 12-AX task.

model weights was trained only on the single fixed mapping of the `12-AX` task, and then tested on the full generalised 12-AX task. While the structure of the model remained identical to the above simulations, i.e. including the loading of the rule - stimulus mapping into the extended set of memory modules, the part of the state space vector containing rule mappings remained constant in all training examples, eliminating the immediate pressure of using the available layer of indirection in training. Thus, the question was whether the model would learn to correctly use the provided mappings during testing, even though during training it did not need to use the mappings at all.

Section 6.2.3 shows that the weight matrices utilising the available layer of indirection through the memory modules conformed to a very specific structure. It was therefore tested if forcing the weight matrix to conform to a multi-diagonal structure would allow the model to achieve rule stimulus separation, even when only trained on a single mapping and compared it to the full bilinear model. This was done by performing gradient descent on a restricted multi-diagonal matrix of the form shown in figure 6.6, instead of the full NxN matrix (or its symmetrised upper triangular).

As shown in figure 6.9, the standard model (labelled “full”) did not immediately achieve abstraction after training on a single mapping. While the model performed errorless as expected on the `12-AX`, on which it was trained, performance after switches in mapping basically resorted back to chance level, despite the memory modules correctly being set to contain the appropriate mapping. Comparing the weight matrices, that were trained upon a single mapping (figure 6.7), with those, trained on a variety of mappings (figure 6.5), clearly showed the stimulus dependent encoding in the learnt weights.

In contrast, models with this restricted weight matrix, despite being only trained on the `12-AX` task, indeed exhibited full generalisation performance, as shown in figure 6.9. They made no new error after switching, provided the mapping memory was updated appropriately. This effect was largely due to the specifics of the regulariser and the chosen encoding, for which the generalised version through the layer of indirection ends up with a lower regulariser cost than the stimulus specific version. Nevertheless, it demonstrated that the system can be set up in such a way that the more general solution is learnt more easily than the stimulus specific version, encouraging abstraction before the first switch.

6.3 Applicability to other models of gated working memory

The sections above presented a possible model for variable substitution based on an adapted bi-linear model. The abstractness and computational capabilities of the bi-linearity is particularly well suited for the implementation of a flexible rule - stimulus mapping through an additional layer of indirection. Nevertheless, the overall solution only depends on the core aspects found in the general class of gated working memory models. As such, similar extensions can also reasonably be expected to work in other models of the class, like the more biologically plausible PBWM. To demonstrate its wider applicability, an adaptation of the LSTM model is now presented to equally include a variable mapping, and show its effects on the generalised 12-AX setup. Furthermore, this experiment return to the setting of sequential online learning, following the same method as the switching and reversal tasks in section 4.8, although the correct stimulus-rule mapping is provided to the network just like in the bilinear model.

6.3.1 Adapting the LSTM for stimulus - rule abstraction

The basic adaptation of the LSTM is equivalent to the one described in the bi-linear model. By enlarging the network with an additional N memory blocks (9 in the case of the 12-AX task), the state space can be extended appropriately to include mappings to abstract the rules from the specifics of the underlying stimulus encoding.

Crucial to the abstraction through indirection is the ability of the network to test for equality / inequality between the content of working memory and the networks input. While the bi-linear model was particularly well suited for this operation, at least in the non-habitual case, the linear weighting of the LSTM cannot do the necessary mapping in a single step. Instead, it draws upon the recurrency of its network. However, unlike for example the PBWM model, it does not have a multi-layer feed-forward or recurrent layer beyond its gated memory blocks. Additional memory units have to be added to the network. In the spirit of shaping, three additional blocks were pre-allocated to act

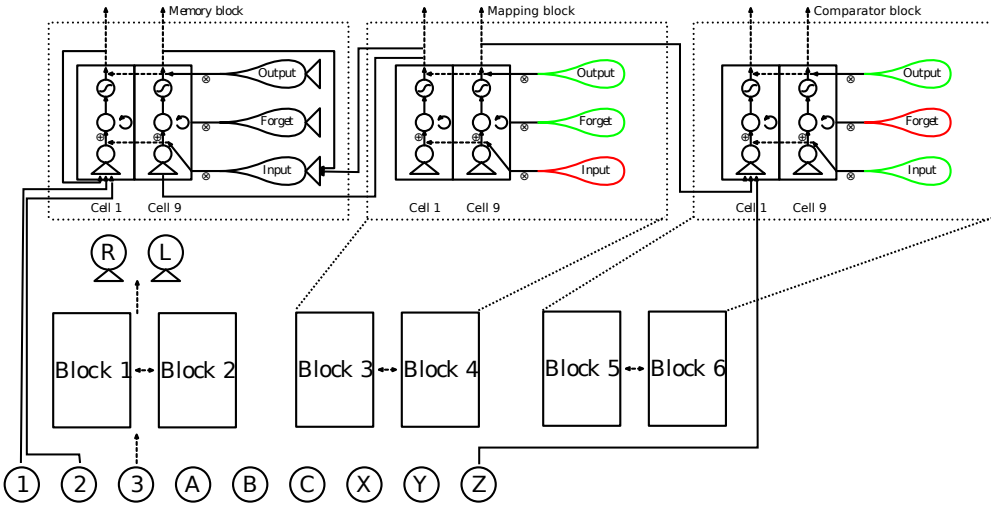


Figure 6.10: Depiction of the adapted LSTM for stimulus - rule abstraction: The standard LSTM network is extended by a number of additional blocks, the mapping blocks and the comparator blocks. Each block has 9 cells included to represent the encoding of one element of the alphabet. For each functional element (2 context markers, 2 x 2 target sequences and 3 distractors) there is one mapping block. The 3 comparator blocks form a two layer feedforward component whose output is available to the rest of the network. Gating is disabled in these blocks with red gates tied to 0 and green gates tied to 1.

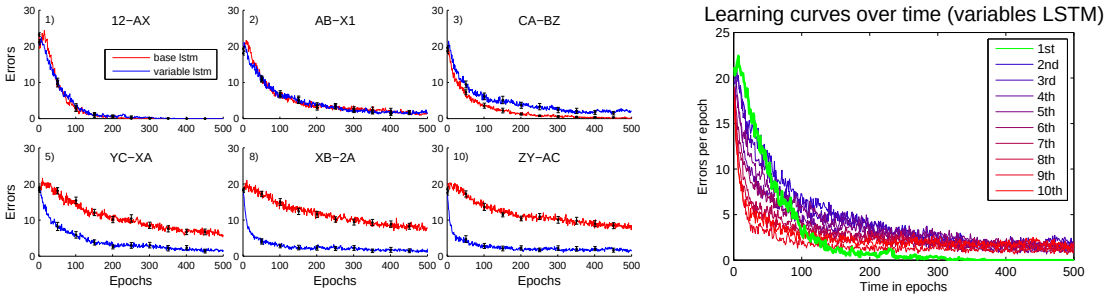
as a generic dedicated comparator network, providing units whose output is 1, if the input vector is equal to the mapping stored in their respective mapping memory blocks. Details of these comparator modules, and how their weights are set, can be found in appendix A.5.

While in previous simulations, the LSTM was chosen to always include 2 cells per gated memory block, this was not sufficient to encode the mapping in an equivalent form to the input vector. As such, for simulations in this section, the network was extended to include 9 cells per block, allowing to store the mapping in a unary encoding, too. Like for the bi-linear model, the actual mapping was provided to the network through an external mechanism. It was updated at the time of switched task in the generalised 12-AX. Overall, the network had 14 memory blocks with 9 cells each, 9 mapping memory blocks, three comparator blocks and two standard memory blocks.

6.3.2 Results

Due to the change of network parameters from 2 to 9 cells per block and from 4 to 2 standard blocks, simulations additionally included standard unshaped LSTM control networks with those parameters.

As shown in figure 6.11, the LSTM network could learn to abstract rule encoding from the stimulus encoding, through the same method of working memory driven layer of indirection as the bi-linear model did. While performance of the network without this



(a) (b)

Figure 6.11: Learning curves of the adapted LSTM on the generalised 12-AX task: (a) shows the progression compared to a non adapted LSTM. Learning performance initially is identical, though with repeated switches the adapted LSTM learns to adjust more rapidly. (b) Learning curves for all 10 switches aligned to the beginning of the switch, showing a strong correlation between number of switches and rapidness of learning.

layer worsened from switch to switch (figure 6.2), showing a positive correlation between the error rate and the number of switches occurred, the opposite was true for the network with the additional mappings, which showed a strong negative correlation (figure 6.11b). With correlation factors around -0.98 in the initial epochs after the switch, this was highly significant and continued to be significant, although with a declining factor, up to about epoch 120. If the first, naive, training was excluded, negative correlation remained significant to epoch 280, after which no correlation remained.

The asymptotic error-rate revealed that the adapted LSTM was not immune to the effect of performance degradation over repeated switches. While 62% of runs performed perfectly within the 500 epochs training time on the 10th switch, 4% of runs no longer were able to learn the task in 500 epochs, with an error-rate of more than 10 errors per epoch. This was in contrast to the naive state, where all runs successfully achieved criterion within 500 epochs.

Unlike the multi-diagonal version of the bi-linear model, the LSTM did not achieve abstraction from the first learning, and only over repeated switches did it learn to achieve this. Even then, only about 10% of runs achieved instant switching after 10 switches. The others required a few epochs to adjust, although much less than from the naive state. There was no specific pressure on the network to favour the stimulus independent solution and switches were sufficiently seldom. The stimulus dependent weights therefore had to be unlearned again at each switch to not interfere with the abstracted ones, and a small amount of adjustment time was to be expected.

6.4 Discussion and conclusion

The ability to vary a set of abstract rules independently from the specific incarnation of particular symbols provides a key component of general cognitive flexibility, and the opportunity to adapt rapidly to an altered environment. This chapter presented an account of how to incorporate this flexibility into the general class of gated working memory models, by providing a layer of indirection between the abstract rule set encoded in weight space and the input layer encoding the environment. It was shown that the class of gated working memory modules provides the architectural components necessary for this abstraction, showing the flexibility of this class of models.

At least two essential components are required to implement the suggested framework, comparable to the those identified by Moody et al. (1998) in their model of DU'S, namely activity-based working memory and a comparator function between the input and the memory layer. Whereas memory is inherent in the whole class of models, the ability to compare and compute with the matched input depends on implementational details. In the bi-linear model comparison is particularly easy, as the multi-diagonal structured matrix provides just that. In comparison, the LSTM needs to draw upon its recurrency, learning a set of weights to achieve the equality matching. As a comparator network is a rather generic function, it is a classic example of a function that might be provided through developmental mechanisms. Alternatively, it may be part of the architectural sophistication necessary, as was the case in the present simulations of the LSTM.

Apart from the networks' ability to realise the abstraction presented here, it was also studied whether it is possible to generalise from the first encounter of the task, or if repeated switches are necessary. While, the LSTM and the standard bi-linear network only generalise once trained upon many variations of the task, it was possible to construct a restricted form of the bi-linear network that generalised immediately. The necessary constraints were rather particular and heavily depended on the implementational details of the model. However, this does indicate the possibility of finding appropriate regularisation.

With the added layer of indirection, the bi-linear model is limited to a class of linearly separable functions. Only through the extension of rules, which provide a form of piecewise linearity, is it possible to capture the full 12-AX task. The degree, to which natural learners and agents indeed are able to provide such flexibility in the more stimulus response driven regime of habits, remains open. There are, however, many ways to overcome the specific limitations of the bi-linear form. Perhaps the least compelling option is the expansion from a bi-linear to a multi-linear or quartic function. Although sufficient, it would drastically increase the number of parameters of the model. The LSTM is able to overcome a similar issue through the use of recurrence. It would be more plausible to change from a localist to a more heterogeneous encoding, as com-

mon in the mammalian brain (Soltesz, 2005; Marder and Goaillard, 2006), and has been shown to play a fundamental role in context-dependent behaviour (Rigotti et al., 2010).

Both models relied upon the correct stimulus rule mapping being already pre-loaded into the appropriate memory elements. While this allowed us to provide an account of a possible substrate of execution, it did not address the important question of how such mappings would be obtained in the first place, and then updated at the time of task switch. One potential solution would be to use the same mechanisms employed in the learning of the PBWM and other models, i.e. to destabilise memory at the time of sustained low reward. However, such random sampling would require considerable time itself, loosing much of the advantages of the flexibility obtained in the first place. Alternatively, a more rule-based goal-directed search may be necessary, resembling the strategy observed in highly trained humans for the generalised 12-AX.

In addition, the strong emphasis on a supervised learning regime of the bi-linear model, requiring specific training examples of the relation between internal state and output, neglected the difficult but important question of on-line sequential learning. For a full account of variable - rule abstraction, it is obviously necessary to address this. There are at least two potential directions through which this may occur.

First, it may be possible to simply include 1:1 mappings for all stimuli initially, reducing the problem back down to the standard learning task. As was shown with the extended LSTM network, a dedicated comparator network can translate this 1:1 mapping into an equivalent representation of the direct inputs of the network. Learning would be no more difficult than in the un-abstracted network, leaving the problem described above of figuring out how to update the mappings in the case of switch.

Alternatively the agent might engage in a self-supervised learning process. With multiple independent learning systems, such as the habitual vs goal directed controller (Daw et al., 2005), one system might train the other, perhaps similar to what is thought to occur during memory consolidation during sleep. Indeed, it has been suggested that sleep is necessary or at least helpful to achieve good generalisation performance in form of "insight" learning. The goal-directed or more flexible learning system is usually thought to learn quicker. But there are studies showing that under some conditions the basal ganglia, typically attributed with the slower habitual learning, reflect learnt behaviour first, before neurons in more prefrontal areas. (Laubach, 2005)

Altogether, it was possible to show that gated working memory models provide a natural extension for the separation of weight based rule memory from specific stimulus encoding, allowing for rapid switches between tasks with related structure. They provide a powerful complement to the sort of flexibility achieved through shaping as described in earlier chapters.

Chapter 7

Conclusions and contributions

This thesis shows the important role multi-task sequential learning can play in achieving flexible cognitive behaviour. It demonstrates with the help of models, methods to overcome limitations that have so far prevented its success in connectionist models. Although the role of sequential learning has long been recognised in various experimental fields (such as behavioural psychology, education, or simply in practical terms of training lab animals to perform behavioural tasks), shaping has not yet taken its deserved place in theoretical neuroscience, particularly in connectionist-style models of learning. There, it has instead been common to focus on more elaborate models.

Such connectionist models have had marked successes over time, providing plausible theories of underlying theoretical principles of cognitive effects (O'Reilly et al., 2010; Rougier et al., 2005), such as the context based switching through excitatory biasing of processing (Cohen et al., 2000; Botvinick, 2007; Herd et al., 2006), the monitoring of conflict and error likelihood (Brown and Braver, 2005), learning of sequential tasks (O'Reilly and Frank, 2005; Kinder, 2000), attention (Pauli and O'Reilly, 2008) or planning (Dehaene and Changeux, 1997). They have also increasingly been able to map these principles and concepts onto anatomical structures, systems and pathways (O'Reilly, 2006; Cohen et al., 2002; Koechlin and Summerfield, 2007; Frank, 2011, 2006; Cohen and Frank, 2009) in the brain, often making some successful predictions about effects of their malfunction (O'Reilly et al., 2002; Frank et al., 2004). Nevertheless, despite their inherent plasticity these models often do not achieve the same rapidness and flexibility of learning as their human counterparts.

Several models of gated working memory were therefore analysed, which contain an additional component of cognitive flexibility, notably the ability to incorporate and draw upon structure learnt in previous tasks, in order to adapt rapidly to the current task at hand.

It was shown that, given the right architectural support, models of gated working memory can benefit greatly from past experience. They could learn complex cognitive tasks

such as the 12-AX task much more rapidly. This is especially the case if the ways by which this experience is gained are deliberately designed by the external environment, e.g. by a teacher, to provide hints and additional information. This is the case for the method of shaping. Even for the moderately complex version of the original parametrisation of the 12-AX task, a setting in which a naive network can still learn, sequential prior knowledge provides an order of magnitude reduction in training time. It was also shown, that the effects of shaping drastically increase with increased complexity of the task. The overall reduction in learning time through shaping is in stark contrast to standard neural network models, which often learn less rapidly when faced with a slowly changing sequence of intermittently unrelated tasks, because of catastrophic interference (French, 1999; McCloskey and Cohen, 1989).

Due to the architectural limitations and requirements shown in both, the supervised and reinforcement learning models, it is clear that sequential learning has to be taken into account when designing models of flexible behaviour. This is especially the case if the model shall benefit from sequential learning, even if each stage is not a strict super set of the previous. Although there are other suggestions how to overcome catastrophic interference at least partially (e.g. Kortge, 1990; Murre, 1992; French, 1994; McClelland et al., 1995), this thesis provided the beginnings of a potential solution in the context of the class of gated working memory. The present solution involves what was termed *resource allocation*. The overall network was modularised into groups of physically separate neural resources. Differential plasticity or learning rates of these groups then leads to a segregation of learnt tasks. This segregation enables learnt weights to be protected from interference. While most of the present simulations employed a form of hard resource allocation, where resources are entirely frozen, a more realistic, soft allocation strategy, in a fully automated form of resource allocation was also suggested.

In addition, it was shown that shaping achieves several additional forms of cognitive flexibility, such as generalisation and abstraction, even if they are not specifically trained for.

By providing one viable solution to the long standing issue of catastrophic interference, this thesis will hopefully help to bring task sequential learning into the focus of neural models of cognition. It shows that the study of sequential learning is viable in neural modelling and can provide an important compliment to the other advances of computational cognition. However, it also has some more practical implications. With connectionist networks beginning to aim to scale-up to ever more complex constructs and tasks, such as tasks of complex visual perception, robotics or sense making, manually setting up network weights in these complex models for any specific tasks is becoming near impossibly difficult. Therefore, training weights through learning algorithms is often the only feasible way, even if one is not interested in learning per se. With the complexity of both task and model increasing, even learning becomes problematic and time consuming when done in one fell swoop. Increasingly (at least in multi

structured models) therefore modellers are applying pre-training algorithms, in which individual components are pre-trained independently of the overall network structure. By drawing upon the ideas presented in this thesis and the understanding of shaping and task sequential learning, some of these network models may benefit from more realistic training methodologies and not needing external intervention. How important a good understanding of staged learning algorithms is, can be seen from the example of deep believe networks (Hinton et al., 2006). Although multi-layer neural networks have been around for a long time and algorithms, like back-propagation, with which to train them equally so, they did not work sufficiently well for many applications of deep networks. Once, however, Hinton et al. (2006) provided a compelling strategy to train multi-layer network in stages, deep belief networks have been applied to various tasks in machine learning. (e.g. Hadsell et al., 2008; Mohamed et al., 2010; Nair and Hinton, 2009)

Nevertheless, shaping is limited in what it can achieve. One important aspect of flexibility that does not benefit from the type of shaping used here, is a form of rule - stimulus abstraction, i.e. the ability to apply the same set of abstract rules to many different sets of stimuli. As typically connectionist networks learn solutions that intrinsically link input stimulus encodings with encodings of the rules in weight space, a change of stimuli often requires relearning of the rules, too. This is particularly the case if the two sets of stimuli overlap in their encoding, such as for example if the meaning of the same set of stimuli is swapped. While for repeated switches back and forth shaping still helps to a limited degree (as long as each of the stimulus mappings is shaped individually), on arbitrary remappings, the simulated networks failed to learn more rapidly on repeated switches.

To account for this type of flexibility in the context of the gated working memory models, a model with the ability to switch instantly between different meanings of stimuli, provided the abstract rules remain the same and the mapping is known to the model, was designed. This flexibility can easily be achieved by introducing a layer of indirection into the model. Instead of encoding rules in terms of stimuli, it lets the rules act upon a set of comparison operations between the input to the model and a group of memory elements. Rules are then specified in terms of whether the input is identical to the content of a specific element of memory, and so a switch in rule - stimulus mapping can be achieved simply by updating the corresponding working memory. It was shown that this method works for other models in the family of gated working memory too, specifically the LSTM. This separation of rules or semantics from stimulus representations may however not only be useful in simple tasks like the generalised 12-AX task, but potentially also in a number of other important cognitive flexibility paradigms. One of these paradigms is “systematicity” (Fodor and Pylyshyn, 1988; Hadley, 1994). This concept refers to an agent’s ability to substitute any object of a given class into a semantic representation of a relation and thus being able to systematically generate new instances of that relation. For example, if a person has

learnt a sentence like "John loves Mary", then they are equally able to replace John and Mary by any other name they know. It is thus a crucial part of the flexibility of language and provides a way to deal with the combinatorial explosion. Another feature of cognitive flexibility is the idea of hypothesis testing. The same form of indirection may be used if representations of "symbols" are extended to include an encoding of rules them selves. This in turn can lead to an instructed general purpose processor that can instantaneously adapt to new tasks, for example through instruction.

Finally, an important part of any theoretical suggestion is experimental validation. In order to augment the computational results of modelling, human behavioural experiments were performed to test the hypothesis that humans would be able to take advantage of hints provided through task sequential learning. It is clear from cases like education that humans under some conditions have the ability to benefit from task sequential learning. Indeed, it is common to deconstruct complex tasks into simpler components that can be taught individually before learning the full task (e.g. Gagné, 1970; Annett, 1996). However, given the real world application, these tasks are usually highly complex, making them less suited for either computational modelling or for detailed understanding of the neuroscientific systems involved. Furthermore, these tasks nearly always contain a strong verbally instructed component, pushing them into a different regime than the purely trial and error driven learning of models. But the degree, to which subjects benefit from task sequential learning in a simple trial and error based learning paradigm, was unclear. The present results showed that humans can indeed take advantage of shaping in this kind of task, confirming the hypothesis and validating the models at least on basic qualitative level. Nevertheless, performance increases were less than expected, leaving the question of whether humans are significantly subject to catastrophic forgetting under these conditions themselves.

Apart from the simple competency-based hierarchical deconstruction in task sequential learning, knowledge transfer can also occur in a number of other, more subtle, ways. One such way may be through the adaptation of representations from one task to the next that can effect efficiency of learning, such as through the categorisation of stimuli. In a variation of the experiment the hypothesis was tested, that subjects would learn the task faster if stimuli were categorised according to the underlying rule set, and thus be able to utilise the additional information provided by the environment through grouping.

Furthermore, a number of additional sequential processes and effects of prior knowledge representations can affect the performance of learning in natural agents. One such effect commonly studied is framing (Tversky and Kahneman, 1981; Levin et al., 1998). It refers to the fact that seemingly inconsequential changes in the formulation of choice problems caused significant shifts of preference. They should be inconsequential as they both provide identical information and thus a rational agent should not make differential decisions. They can effect the weighting of various properties (such as risk preference)

in evaluating the value of decisions. Another cognitive phenomenon that can effect learning is confirmation bias (Nickerson, 1998; Jones and Sugden, 2001), which results in a tendency, when testing an existing belief, to search for evidence which could confirm that belief, rather than for evidence which could disconfirm it. Although the influence of either effect on subjects performance in the 12-AX task cannot be ruled out, it does seem unlikely though. Given that stimulus presentations in the task are not contingent on the choice of responses, subjects do not actually have the option to seek out specific information and are predominantly passively consuming the information. Also, given the deterministic (non-probabilistic) nature of the rules in the 12-AX task, a selective over or under-weighting of evidence is difficult, as there is a binary outcome of either being consistent or not. To show confirmation bias, subjects would need to outright ignore some stimuli inconsistent with their hypothesis they currently maintain. Framing effects are similarly unlikely, again due to the deterministic nature of the rules. One task sequential effect that is perhaps more likely to occur is that of motivational influence. Especially in the more demanding learning condition of the Apple-Axe task, subjects may need to continue to learn a single set of rules over the duration of more than an hour without giving up. Although there are many factors that influence motivation and resignation, one such task sequential influence is the question of how much control over success a subject might have had in previous experiments, as has been demonstrated in the case of learnt helplessness (Maier and Watkins, 2005; Maier, 1984)

In addition to testing the hypotheses that humans can benefit from task sequential learning and prior knowledge of a cognitive rule based task in a trial and error based learning paradigm, the experiments also provided a baseline learning performance for the basic `12-AX` task. As this task has to date been used in a number of computational studies (e.g. Frank et al., 2001a; Hazy et al., 2007; O'Reilly and Frank, 2005; Zilli and Hasselmo, 2008; Dayan, 2007; Todd et al., 2009), yet no behavioural data existed, these results allowed to finally identify how well these computational models capture human performance on this task.

Data for several parametric variations was provided, which proved the task to be sufficiently rich as to scale between being easily learnable in a single session to being barely learnable without additional time or help, such as shaping. It was thus shown that this task is interesting and revealing for human behavioural studies too, lending itself well in understanding the effects and relevance of gated working memory. In the future, it may even play a wider role, joining other prefrontal dependent tasks in behavioural and neuroscientific characterisation. It has already been suggested as a candidate task to evaluate performance of "rule generation and selection" in clinical trials (Barch et al., 2009b), a key component of executive function thought to be impaired in schizophrenia, although it was considered not fit yet due to lack of experimental data..

While the present results showed that the models presented still perform worse than humans, this thesis provides insight into the role, the benefits and the difficulties of

sequential learning to accomplish cognitive flexibility. It highlights the need to consider sequential learning in the endeavour to understand and model the impressive cognitive flexibility of rapid adaptability of natural learners. But it will be important in future to extend this work and link it more closely with the various other methods used in studying sequential learning to create a fully integrated understanding.

From a theoretical point, the mechanisms of structural segregation, or what was called resource allocation, will continue to need further investigation and experimental verification. One of the key questions to explore will be the mechanisms by which any such form of structural assignment of learning may occur and in what form it is triggered.

In the past, there has been a variety of models employing structural segregation for learning such as the Mixture of Experts model (Jacobs et al., 1991a), MOSAIC (Haruno et al., 2001), compositional Q-Learning (Singh, 1992) or ART (Carpenter and Grossberg, 1988). Common to all of these is the idea to use a combination of an actor or forward model and an inverse, predictive model. Each modular structure has its own predictive model with a global gating network arbitrating between modules. Learning from individual samples then occurs in each model proportional to its predictive quality of the outcome. Although at least MOSAIC and CQ-L also operate in a temporally extended fashion, these models distinguish themselves from the setting considered in this thesis in that they learn in a stationary environment, i.e samples are drawn randomly and repeatedly from all the conditions covered by the different modules. In contrast, in the setting of cognitive shaping explored here, tasks were not explicitly revisited once they had been learnt, instead following a continuous progression in tasks. Thus, this does not allow for a gating network to gradually carve out the input space to which each module should respond. Furthermore, due to the hierarchical nature of the tasks considered, individual modules were not fully independent of each other, as they together formed the state representation upon which each forward model acted. Therefore the present allocation followed a different mechanism in which a global event triggered a stepwise shift in plasticity, mitigating the issue of non-revisiting while still allowing for modularity. This is not entirely dissimilar to the mechanism of Redish et al. (2007) for modelling simple extinction and renewal learning, in which a split in state space occurs at the boundary from conditioning to extinction. This state space expansion is triggered by a reduction in expected reward, just like in the current simulations.

In gated working memory models, the actions themselves determine the state space representation. Thus, with adaptation or learning of the internal action selection, the state representations change. Due to the hierarchical nature of the task, modules are not fully independent as they may draw upon the full state space if needed. Here, the continuous growth allocation might have an advantage, as learning in prior modules is disabled or reduced and thus provides a stable platform to build upon. There have been suggestions in the context of hippocampal cortical interactions, however, of how to maintain episodic memory in the face of cortical semantic plasticity (Káli and Dayan,

2004), which could potentially alleviate the need for stable representations.

One disadvantage of a continuous growth allocation is the question of reuse and adaptation. Unlike in the predictive approach, later observations can no longer adjust learnt weights, even if they come from the same or similar tasks. Indeed, this disadvantage resulted in requiring a small but finite base plasticity in all models for automatic allocation to be successful in the current simulations with the LSTM network. Nevertheless, the lockout of adapting older tasks may also have advantages, in that related tasks do not interfere, allowing to retain a number of similar tasks. Indeed, probably the closest observed phenomenon to such an allocation method is the adult neurogenesis of the hippocampus (Deng et al., 2010), which has been suggested to play an important role in reducing catastrophic interference by reducing overlap in the sparse code of dentate gyrus (Aimone et al., 2010; Wiskott et al., 2006) and is thought to occur more strongly in the situation of learning (Gould et al., 1999).

A further open question, resulting from the event-triggered expansion of the network, is that of capacity. So far, no constraint on the growth of the network were included. While again the responsibilities-based gating allows to scale the available modules naturally to tile the space optimally, this is not the case for the event-triggered expansion, which would continue to grow without bound. So far, it is unclear what capacity constraints might exist, particularly as the models have directly linked storage capacity with processing or transformational complexity, as can be seen in the slightly odd use of memory cells for the comparator network used in the stimulus-rule abstraction. While Frank et al. (2001a) have suggested there might be as much as 20,000 stripes in PFC which they have linked with the modules of gated working memory, the capacity of explicit central working memory store is thought to be as low as 4 - 7 items (Cowan, 2001; Miller, 1956). The question of how to recycle modules once capacity is reached and its implications of different recycling strategies on learning are important to explore.

Thus, it will be important to explore the limitations and advantages of this allocation strategy more thoroughly. An interesting question is if it can be experimentally determined which approach, event-triggered or responsibility-based, matches behavioural data more closely. However, it is likely also interesting to try and combine the two approaches. In a model designed for event segregation Reynolds et al. (2007) have for example used a global predictive model to detect boundaries of increased unpredictability to trigger an update in its internal representations, but did not use it to drive modularity.

Beneath the layer of structural modularisation, encoding plays the role of determining representations. While the present representations have mostly been localist, neural encoding is commonly seen as distributed (O'Reilly, 1998; Rumelhart et al., 1986), although the case has been made for the biological plausibility of localist representations, too (Bowers, 2009). However, this is not an important element of the models, and binary rather than unary codes have been used in simulations, too. Furthermore,

Reynolds and O'Reilly (2009) have shown that gated working memory models can utilise and learn distributed codings. However in this model too, it remains that a hard separation between individual memory modules occurs with no crosstalk beyond a unit of gating resulting in localist representations on the modular level. Nevertheless, PFC neurons are known for their diverse response properties. For example, Rigotti et al. (2010) have suggested that this mixed selectivity is important to enable this cortical area to sub-serve flexible cognitive behaviour through the complex conjunctions of events and inner mental states. Thus, one interesting question would be whether gated working memory modules can further be distributed to include mixed selectivity across modules of gating. Particularly important would be to equally extend resource allocation, or differential plasticity, into the setting where no neuron is associated with a unique module.

So far, the present models have been associonist, and mostly stimulus driven. Although as with any cognitive model, manipulations of internal state space are critical to the models operation, any updates were directly triggered by external stimuli. The models do not have or exhibit the ability of stimulus independent thought or forward planning, which would allow them to make sophisticated predictions. These models can be seen as model-free, despite flexible behaviour of the prefrontal cortex often being associated with model-based planning. Although the behavioural experiments could neither confirm nor reject the question of whether human behaviour and learning in the 12-AX task is driven by either a model-free or a model-based approach, trying to extend the class of gated working memory models by a hypothesis-driven model would be a valuable addition. Perhaps the easiest route to study hypothesis-driven learning is to consider the generalised 12-AX task or WCST. Here, once the base task is learnt, the hypothesis space is clear as it is limited to the possible rule - stimulus mappings. With a model like ours, testing a new hypothesis would only require gating new mappings into working memory. One shortcoming of the models used here is that they do not utilise feedback or reward in the model itself, but only use it for learning. In order to trigger gating and therefore change to a new hypothesis, the reward or error signal is needed, though. It would, however, be easy to incorporate reward as an input signal directly, and indeed, in the PBWM model unexpected lack of reward can trigger gating actions itself.

Another interesting question would be if the form of rule retrieval suggested by Dayan (2007) could also be used in hypothesis driven learning. If sets of rules were grouped together, for example through a task sequential allocation mechanism, it should be possible to associatively match on a part of state space, representing the hypothesis to be tested. Then the same mechanism of gating new hypothesis into state space could be used as above. How learning such a mechanism might work, in order not to have to refer back to a homuncular element, remains yet to be explored.

The kind of tasks considered are inherently hierarchical in nature, a feature common

in both cognitive and other tasks. This hierarchy exists particularly on a temporal stability of state representations or memory storage, and is characterised by nested sets of contextual representations influencing action. For example in the 12-AX task, three levels exist with the stimulus, the one-back and the contextual level. It is possible for example to link these hierarchical levels of the task to the hierarchical levels of cognitive control in the cascade model of Koechlin et al. (2003); Koechlin and Summerfield (2007), in which the task would most likely require the roles of the stimulus, contextual and episodic control.

The situation of task sequential learning extends this hierarchy one level higher, with lower levels conditioned on the task at hand, allowing rapid and flexible adaptation on an even longer time scale. The concept of gated working memory also fits well with this notion and particular with the concept of resource allocation. Just like the operations of gating that trade-off between stability and plasticity of a unit of working memory, resource allocation controls the same trade-off on the encoding of task rules. However, while gating acts on activity-based and stimulus-driven working memory, resource allocation acts on weight-based learning. Rules are commonly thought to be encoded in weight space and are modelled that way here. This difference is, however, not necessary. If rules were encoded in an activity-based way it should increase the similarity even further. For example, in a hypothesis-driven approach as suggested above, rules could be encoded through recurrent activity, and indeed one can find activity of neurons correlating with rule identity in PFC (e.g. Hoshi et al., 1998; Watanabe, 1990).

Some of the previous accounts of computational shaping (e.g. Elman, 1993) have considered even longer time scales of developmental proportion. While conceptually it is quite similar to the shorter task sequential setting, the mechanism of modularisation or resource allocation is quite different. Instead of the network itself driving the structural plasticity based on events in the sequential stream of inputs, it occurs independently over time, although it might to a limited degree be triggered by events, too (de Villers-Sidani et al., 2008). Therefore, either the timing of the shaping schedule has to be adapted to fit into the correct development period of the brain, or the capacity expansion of the network itself drives the shaping through differential interpretation of the identical external task, leading to a form of self-shaping (Elman, 1993).

On a similar temporal extension over the lifetime of the agent, spanning a multitude of tasks, is the concept of lifelong learning (e.g. Thrun and Mitchell, 1995). Unlike shaping though, it isn't focused on learning an individual task, instead emphasising learning to learn. Rather than constructing individual skills and competencies upon which later tasks can build, it's aim is to learn a bias or constraint to enable rapid learning from only a few samples. By constructing a prior over hypothesis (Baxter, 1997), through lifelong experience, learning can be much more rapid. As an example in the structural domain (Braun et al., 2010), Kemp and Tenenbaum (2008) showed that it is possible over the time course of many examples to learn a probabilistic relational structure that

can then be used to more rapidly learn properties of new objects by inference from nearby known objects. Transferred to the setting of cognitive tasks, this would allow over a long developmental time-scale to learn a prior distribution over the class of all tasks, that can later be used in efficient inference of the task rules of a yet unseen task. It would be interesting to extend this model to cover not only task sequential learning but this form of learning to learn, too. It is likely that learning appropriate representations from multiple tasks (Rougier et al., 2005) will play an important role in this, together with perhaps a rule based mechanism like in the bi-linear model.

On the other side of the hierarchy is the stimulus - rule mapping, used for e.g. the generalised 12-AX task which abstracts away the concrete representation of the stimulus and inserts a layer at the bottom of the hierarchy, although the frequency of updating these corresponds more with the highest or task level.

As models of natural cognition and learning, however, not only theoretical questions need to be taken into account, but also experimental, making it increasingly important to link these ideas into specific experimental validation.

So far, there is only limited direct biological evidence for the form of resource allocation suggested. This caveat is specially true for the hard resource allocation that was generally used throughout these simulations. However, many of the basic ingredients necessary for such a system do exist in the brain, such as meta plasticity via the effects of neuromodulators on learning. It will be important to identify and study potential biological and systems foundations of modularising processing at task boundaries. While gating of working memory has been suggested to occur through cortical basal ganglia interactions, it yet has to be made clear which area of the brain might sub-serve the control and regulation of meta plasticity. It would also be interesting to identify the degree to which anterior PFC is involved in switches of task. While no working memory or state needs to be kept across task boundaries, leading to switching rather than branching (Koechlin et al., 1999), the need to combine different tasks to transfer skills between them, together with the explicit instruction in the set of experiments to keep old tasks in mind, might still engage aPFC. Furthermore, this would raise the question if the limit to hold more than one task concurrently (Charron and Koechlin, 2010) may have affected the effectiveness of the shaping protocol to accelerate learning the 12-AX. As, however, even the base 12-AX task without task level switching activated aPFC (Reynolds, 2007), it may be difficult to distinguish.

Even the model of basal ganglia being responsible for gating working memory itself is only beginning to be explored and experimentally validated. As it was shown here that the 12-AX task is robustly learnable, depending on the variant, it would thus be interesting to study its learning in fMRI. For one, it would be interesting to see if differences between hypothesis driven learning and statistical associative learning can be identified and if there is an interaction between the two. Also the differences between learning and steady state execution of the task may reveal interesting facets

about rule learning. But even from a purely behavioural level, there are still a number of interesting options to be explored about the 12-AX task, especially such as trying to verify some of the predictions made with respect to shaping and generalisation.

Finally, with a greater understanding of the algorithmic and implementational level of task sequential learning, it may in future also be possible to predict more accurately which shaping sequences might be most efficient for a specific task. So far, the shaping sequence was defined somewhat arbitrarily, falling into the tradition of shaping being "more like an art than a science" (Midgley et al., 1989). Nevertheless, it did follow a principle akin of hierarchical task decomposition (Gagné, 1962), but the behavioural performance shows that there is still room for improvement. Furthermore, with increasing complexity of the task and thus the number of possible shaping protocols, the difficulty and importance in choosing an optimal protocol grows. Therefore, when applying the design for optimal task sequences in real world applications, like education, a thorough understanding will be needed. While for example Resnick et al. (1973) did use a methodical approach in constructing a curriculum for teaching mathematics, it is a purely psychological approach and takes neither specific computational considerations, nor results from neuroscience into account, which is increasingly being attempted (Goswami, 2006; Goswami and Szucs, 2010). Hence in the long run, ideas from this thesis together with further work might contribute to the transfer from neuroscientific and modelling research to educational practise.

Appendix A

Additional material and equations

A.1 Stimuli presented to subjects

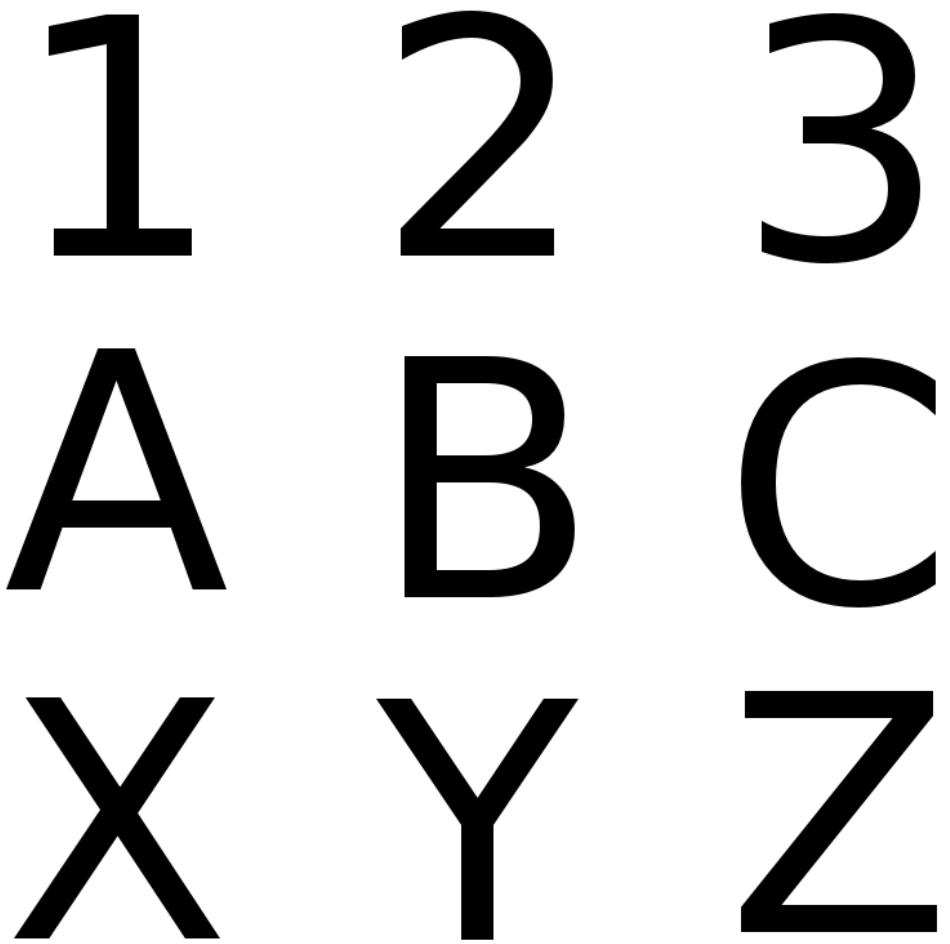


Figure A.1: Stimuli used in the 12-AX task

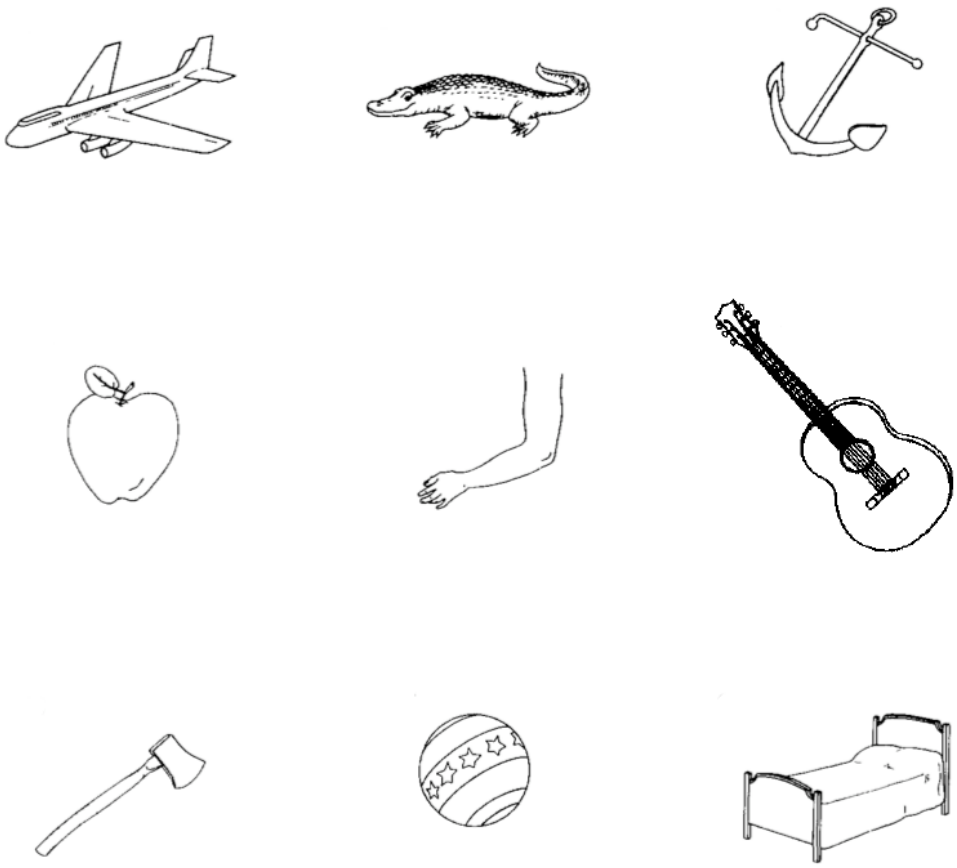


Figure A.2: Stimuli for the Apple-Axe task

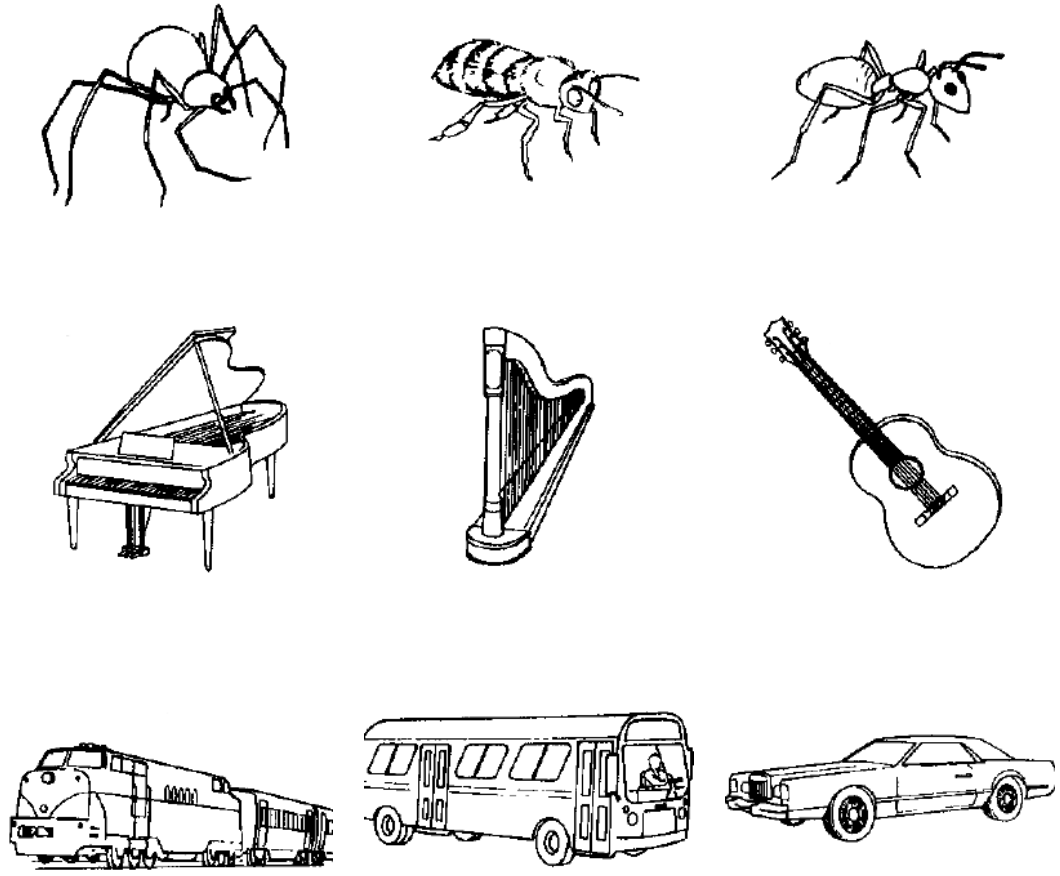


Figure A.3: Stimuli for the Spider-Train task

A.2 Instructions given to subjects

Information Sheet for Participants in Research Studies

You will be given a copy of this information sheet.

Title of Project: **Learning of rules for sequential responses across tasks**
This study has been approved by the UCL Research Ethics Committee 2005/001

Prof. Peter Dayan (Principal Researcher)
+44 (0) 20 7679 1175
Gatsby Computational Neuroscience Unit
17 Queen Square
London WC1N 3AR

Kai Krueger
+44 (0) 20 7679 1172

We would like to invite you to participate in this research project. You should only participate if you want to; choosing not to take part will not disadvantage you in any way. Before you decide whether you want to take part, it is important for you to read the following information carefully and discuss it with others if you wish. Ask us if there is anything that is not clear or you would like more information

We are interested in how people learn and figure out simple rules of responding to sequences of characters based on feedback.

The task you will participate in is fully computer-based. You will sit in front of a computer and observe characters, one at a time, on the screen to each of which you need to respond with one of two buttons (or) according to some rules. Your goal is to figure out these rules. The task is split into two phases, a learning phase and a test phase. After each of your responses (in the learning phase), you will receive feedback to say if your response was correct. In case of a correct response, you will see a green, happy, smiley face; if it was incorrect you will see a red, sad, face. Since one of the two possible responses is always correct, if your choice was incorrect, pressing the other button would have been correct. Once you have received feedback as to the correct response, the next element of the sequence will be presented, and you will again need to respond. You can go at your own pace, but there is a timeout after 20 seconds (indicated by a grey neutral face), after which it continues with the next element of the sequence.

The correct response to each element of the sequence is determined by a set of rules. Your task is to try and learn these set of rules and receive as much correct

feedback as possible.

Although we can't tell you the exact rules, as this is what you are trying to learn, we can tell you that the rules have the following properties:

- The rules and feedback are deterministic and reliable, i.e. if you give the correct response you will always get positive feedback and the correct response is fully determined by the sequence you have seen so far. Therefore, once you have figured out the rules it is possible to get 100% correct feedback.
- The rules and thus your response depend on the current and some of the previously presented characters. Therefore you will need some memory, although you don't have to remember long sequences.
- Neither your response, nor the correct response, plays a role in the rules and therefore they don't influence how you need to respond to future characters. They are just presented as feedback. That is, the correct response is solely determined by the presented characters.

Although the rules aren't complicated, this way of learning them (by trial and error) can be difficult. It is therefore completely normal that it may take a long time for you to learn them (of the order of an hour or two). You will thus be encouraged to just continue to try and learn the rules fully even if it may well feel a bit frustrating continuing to make errors. There is a good chance eventually it will click and you realise what the rules are. But even then not everyone succeeds in learning them, so try your best, but don't worry if you don't succeed fully.

The task is divided into 'epochs of about 200 responses. After each epoch, you will have a moment to relax, stretch and make sure you feel comfortable to continue. This break is clearly indicated on the screen and you can resume with the next epoch whenever you feel you are ready to continue.

For some participants there will be several switches of rule sets, each of which will be clearly indicated. In all cases though the rule set will remain constant unless explicitly indicated. The various rule sets have some commonality that may facilitate learning and all obey the properties outlined above.

The first (training) phase will complete either once you have learnt the rules well enough (90% correct in two consecutive epochs) or after a maximum of 25 epochs, should you not reach the criterion beforehand. At this point, you will automatically progress into the second phase, the testing phase, again clearly indicated. This is very similar to the previous phase, with the identical task and rules, just that you wont get any feedback if you are right or wrong until the end. It is also shorter with only 2 epochs.

You will receive 7.50 per hour to offset any expenses you might incur.*Also in*

addition, to reward you for your performance, you will receive a bonus (up to 7.50), depending on how well and quickly you have learnt the task. This is assessed by the time it takes you to reach the learning criterion (90in two consecutive epochs) and your performance during the test trials. The harder to learn, but less frequent responses are rewarded more.

Participating in this study will typically take approximately one hour to complete. The experiment will not last longer than two hours in any case.

To participate, you must have normal or corrected-to-normal vision and should not suffer of any learning disabilities.

It is up to you to decide whether or not to take part. If you choose not to participate it will involve no penalty or loss of benefits to which you are otherwise entitled. If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form. If you decide to take part you are still free to withdraw at any time and without giving a reason.

All data will be collected and stored in accordance with the Data Protection Act 1998.

A.3 Raw data graphs of subjects

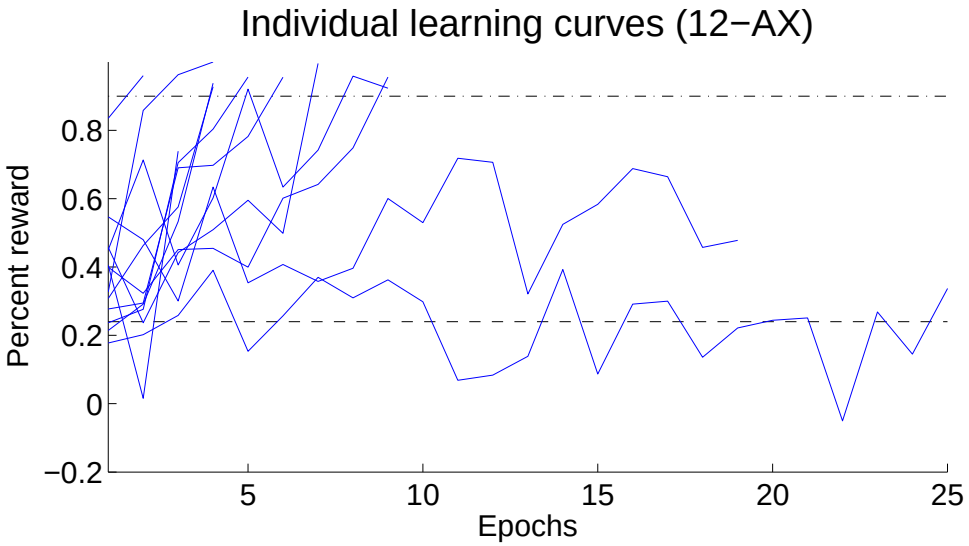


Figure A.4: Individual learning curves for all subjects of the 12-AX task

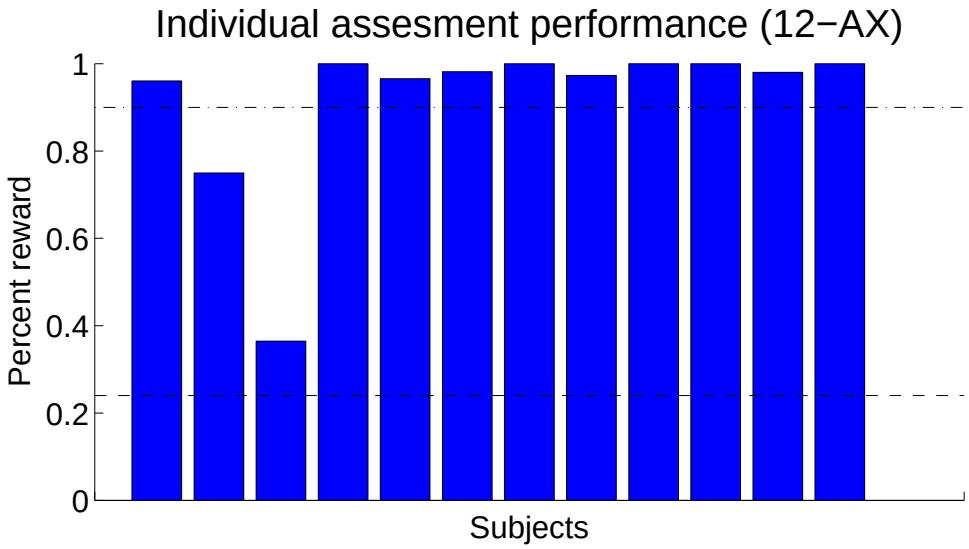


Figure A.5: Individual performance during assessment for all subjects of the 12-AX task

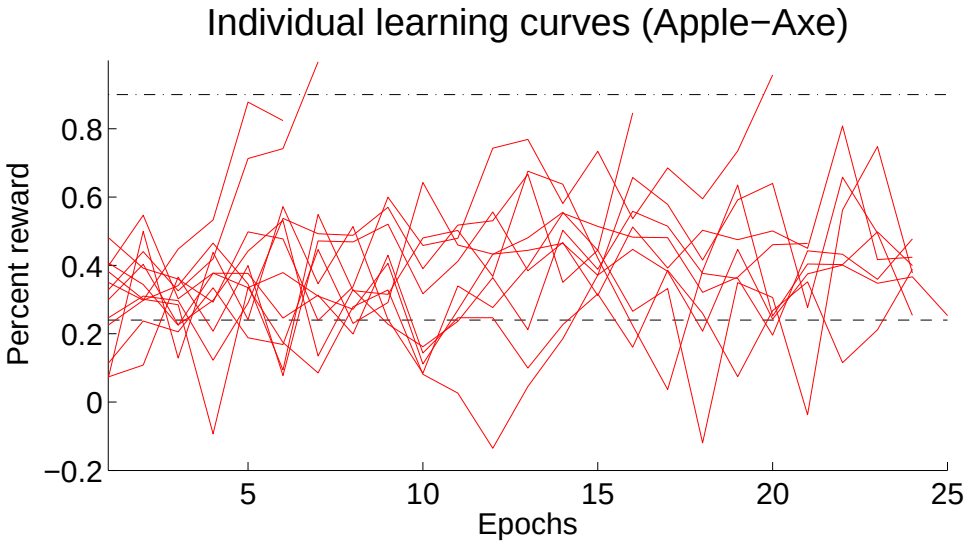


Figure A.6: Individual learning curves for all subjects of the Apple-Axe task

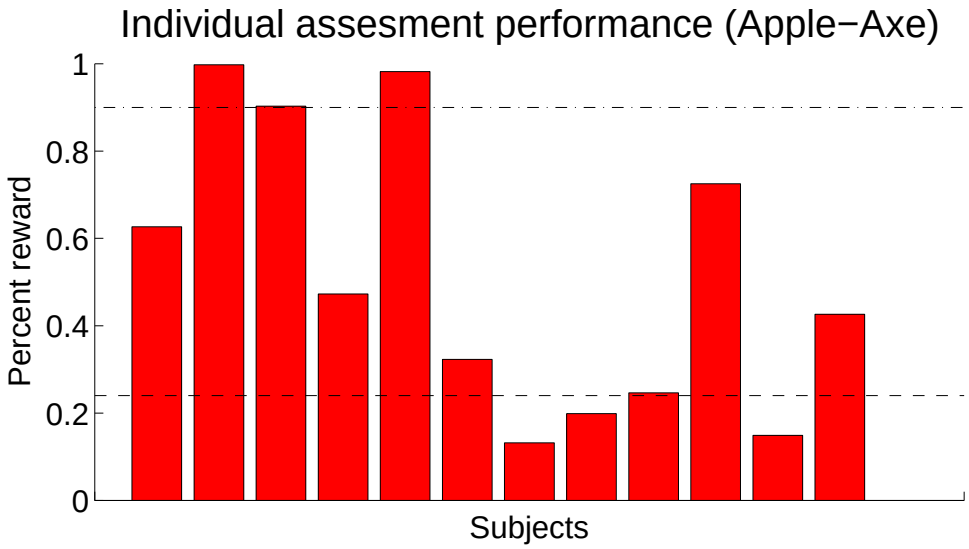


Figure A.7: Individual performance during assessment for all subjects of the Apple-Axe task

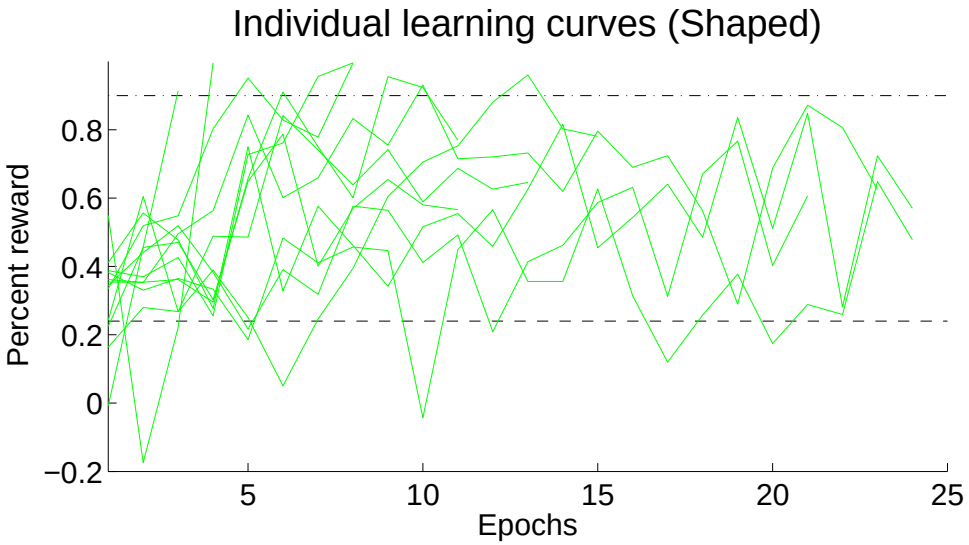


Figure A.8: Individual learning curves for all subjects of the shaped Apple-Axe task

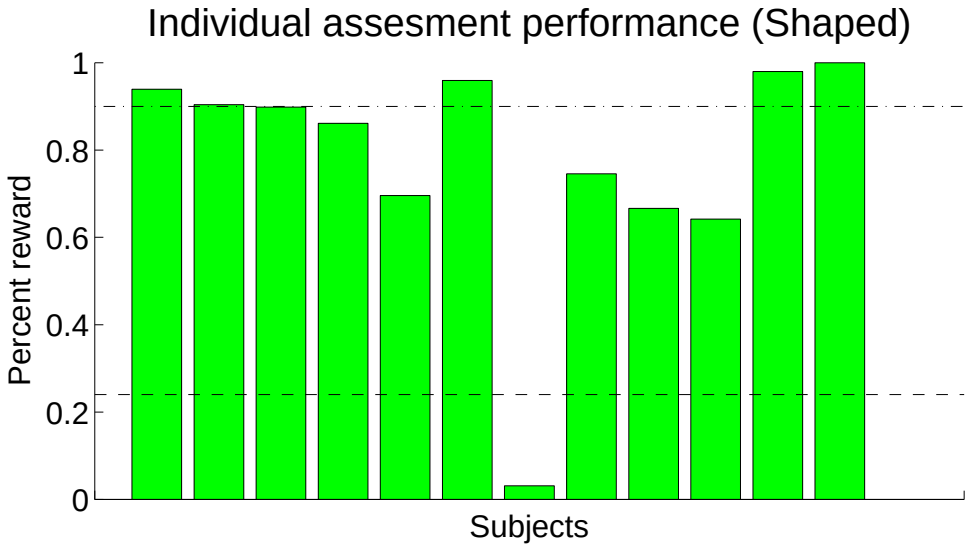


Figure A.9: Individual performance during assessment for all subjects of the shaped Apple-Axe task

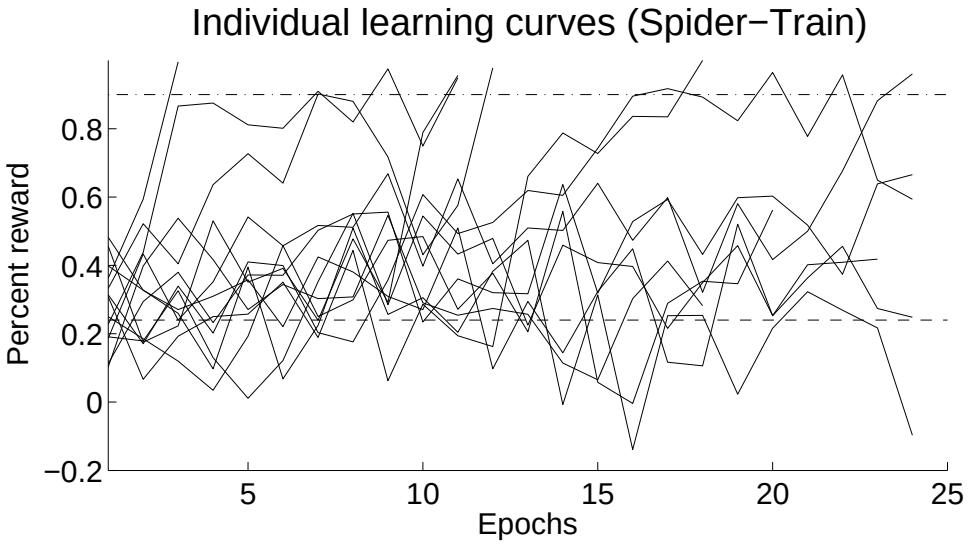


Figure A.10: Individual learning curves for all subjects of the Spider-Train task

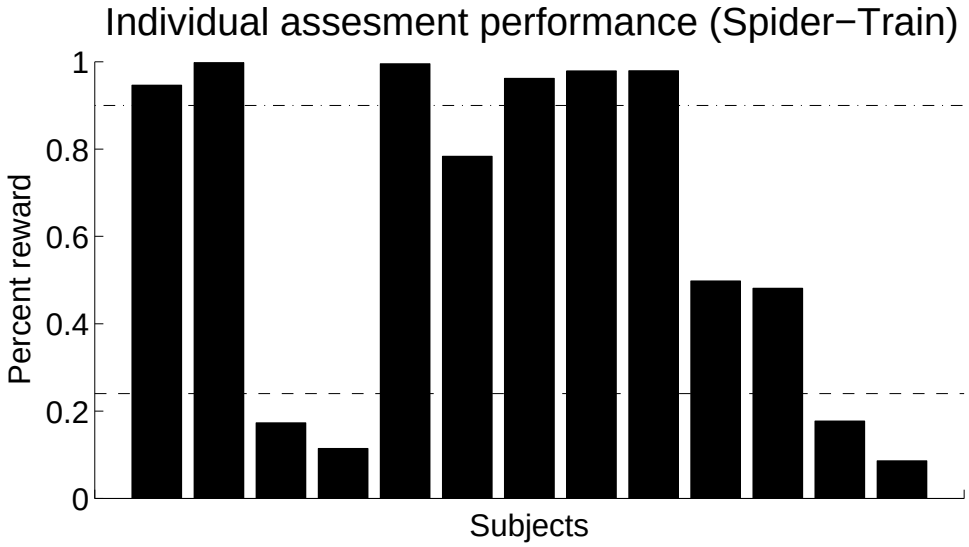
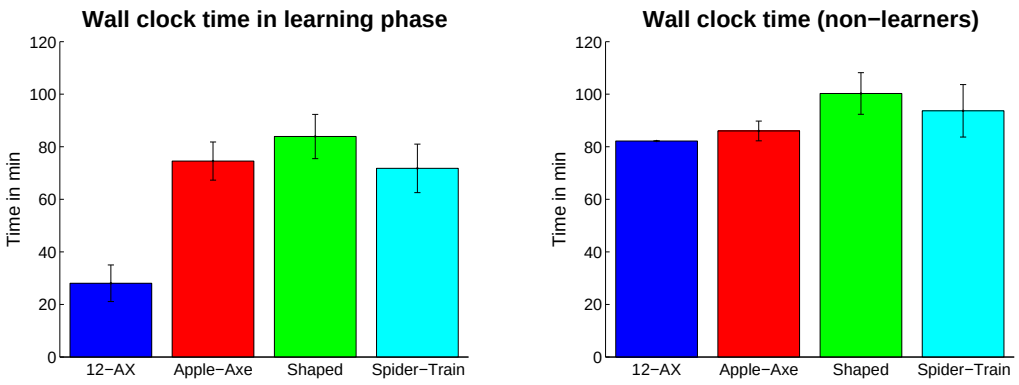


Figure A.11: Individual performance during assessment for all subjects of the Spider-Train task



(a) (b)

Figure A.12: Average wall clock time subjects spent in the learning phase (time from beginning of the first train epoch to the beginning of the first assessment epoch). This time includes any breaks subjects took in between epochs and also includes the time in the shaping tasks. (a) Times averaged over all subjects independent of their success of learning. (b) Average over subjects who did not achieve the learning criterion in the full task.

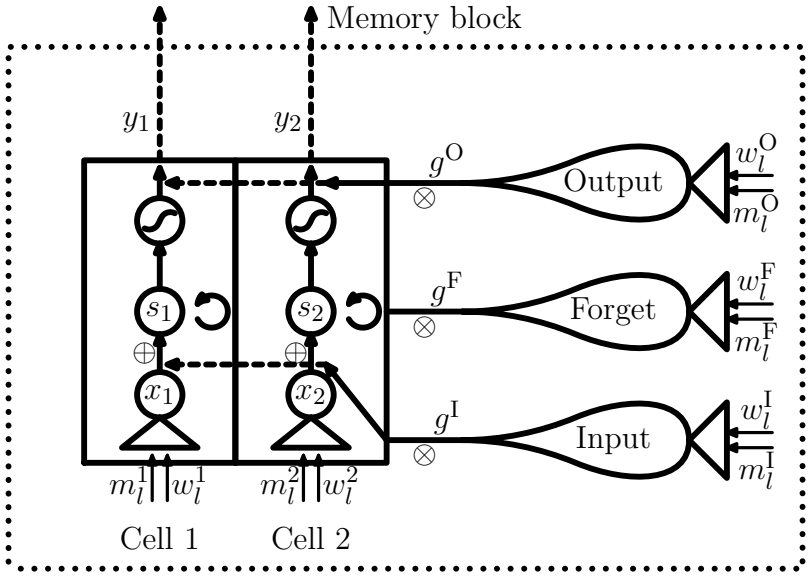


Figure A.13: A single memory block of the LSTM together with the notation used in the equations

A.4 Network equations for the LSTM

A.4.1 Forward pass

Here we present the detailed equations of the forward pass of the LSTM, i.e. the propagation of activations through the network, before presenting equations involved in the backward pass or learning stage in section A.4.2. These equations come from the original paper describing LSTM with forget gates by Gers et al. (2000) and are the ones used in our simulations in chapter 4.

The LSTM network consists of five types of units: input units with unary (0/1) activations i_l based on the stimulus presented; output units o_R, o_L, indicating which choice the network makes; and then each memory block (see figure A.13) has three different sorts of gating unit $g^{\mathrm{I}}, g^{\mathrm{F}}, g^{\mathrm{O}}$, and three units for each cell c within a block, input x_c, memory s_c and output y_c. Sigmoid and tanh functions are used as activation functions throughout, based on linearly weighted inputs. The activity of the memory cells at one time step acts as inputs to all the relevant units at the next time step. We describe the equations for a single memory block below. When the network contains multiple memory blocks, all the units in each block receive inputs from all the other blocks.

Gating units: Gating term $g^{\mathrm{I}}(t)$ is calculated as:

$$g^{\mathrm{I}}(t) = \sigma\left(\sum_l m_l^{\mathrm{I}} y_l(t-1) + \sum_l w_l^{\mathrm{I}} i_l(t)\right) \tag{A.1}$$

where $\sigma(z) = 1/(1 + \exp(-\beta z))$ is a standard logistic sigmoid function, m_l^I are recurrent weights from the memory cells y_l, and w_l^I are feed-forward weights from the input units i_l. $g^{\mathrm{F}}(t)$ and $g^{\mathrm{O}}(t)$ are calculated similarly. β was chosen to be 1.5.

Memory input units: The input to memory cell c is

$$x_c(t) = 2 \tanh \left(\sum_l m_l^c y_l(t-1) + \sum_l w_l^c i_l(t) \right) \tag{A.2}$$

Storage unit: The stored content of memory cell c at time t ($s_c(t)$) is calculated as its content at the previous time step combined with the gated input. At the beginning of each epoch, the memory cells' contents are reset to zero $s_c(0) = 0$

$$s_c(t) = s_c(t-1) \times g^{\mathrm{F}}(t) + x_k(t) \times g^{\mathrm{I}}(t) \tag{A.3}$$

Memory output unit: The output depends on another non-linearity

$$y_c(t) = \tanh\left(s_c(t)\right) \times g^{\mathrm{O}}(t) \tag{A.4}$$

Choice units: Finally, the choice of the network is based on the two output units o_L and o_R, which are calculated as

$$o_R(t) = \sigma \left(\sum_l m_l^R y_l(t-1) + \sum_l w_l^R i_l(t) \right) \tag{A.5}$$

Although this form of output is used for learning, in evaluating the performance of the network, we binarise the unit's output, equivalent to setting $\beta = \infty$. Activations of $[1, 0]$ correspond to a left response, $[0, 1]$ to a right response, and both $[0, 0]$ and $[1, 1]$ are considered invalid responses, and are counted as errors.

A.4.2 Learning in the LSTM model

The function being minimised is the usual square error loss function between the target output and the actual output

$$E(t) = \frac{1}{2} \sum_{k=\{R,L\}} \left(t_k(t) - o_k(t)\right)^2 \tag{A.6}$$

This is minimised by gradient decent, using the update equations for each of the weights

$$\Delta w_l^m(t) = -\alpha \frac{\partial E(t)}{\partial w_l^m} = -\alpha \sum_k \frac{\partial E(t)}{\partial o_k(t)} \frac{\partial o_k(t)}{\partial w_l^m} = \alpha \sum_k e_k(t) \frac{\partial o_k(t)}{\partial w_l^m} \quad \text{(A.7)}$$

$$= \alpha \sum_k e_k(t) \frac{\partial o_k(t)}{\partial y_l(t)} \frac{\partial y_l(t)}{\partial net_l(t)} \underbrace{\frac{\partial net_l(t)}{\partial w_l^m}}_{y_m(t-1)} \quad \text{(A.8)}$$

$$\Delta w_l^m(t) = \alpha \frac{\partial y_l(t)}{\partial net_l(t)} \left(\sum_k e_k(t) \frac{\partial o_k(t)}{\partial y_l(t)} \right) y_m(t-1) \quad \text{(A.9)}$$

with $e_k(t) = t_k(t) - o_k(t)$

Calculating $\frac{\partial o_k(t))}{\partial y_l(t))}$ however depends on which type of unit y_l is and is described separately for each.

Output units: In the case of the output units, $\frac{\partial o_k}{\partial y_l}$ becomes trivial, as it is 0 for $l \neq k$ and 1 for $l = k$, thus leaving

$$\Delta w_m^l(t) = \alpha e_l(t) \frac{\partial o_l(t)}{\partial net_l(t)} y_m(t-1) = \alpha e_l(t) f'(net_l(t)) y_m(t-1) \quad \text{(A.10)}$$

This is the standard back-propagation equation.

Output gating units: The first layer of hidden units with respect to the output layer are the output gating units. Although they are still on the output side of the network, they are connected to the output units via the multiplicative attenuation of the memory cell output. Furthermore, due to the recurrent nature of the network, the memory output unit, and thus the activation of the output gating, feeds back into the memory cells. However, as described in Hochreiter and Schmidhuber (1997), this error flow is truncated once it leaves the memory blocks and only the direct error flow is considered in the learning equations. In their empirical evaluation this truncation does not appear to affect the network's ability to learn long time lags. With multiple memory cells per memory block, each output gating unit affects all memory output units of a given memory block:

$$\frac{\partial o_k(t)}{\partial y_l(t)} = \frac{\partial o_k(t)}{\partial net_{o_k}(t)} \frac{\partial net_{o_k}(t)}{\partial y_l(t))} \quad \text{(A.11)}$$

$$= f'(net_{o_k}(t)) \frac{\partial \sum_c w_k^c y_c(t) y_l(t)}{y_l(t)} \quad \text{(A.12)}$$

$$= f'(net_{o_k}(t)) \sum_c w_k^c y_c(t) \quad \text{(A.13)}$$

This leaves as the update equation

$$\Delta w_l^m(t) = \alpha f'(net_l(t)) \left(\sum_k e_k(t) f'(net_{o_k}(t)) \sum_c w_k^c y_c(t) \right) y_m(t-1) \tag{A.14}$$

Hidden layers: For the remaining units, the error flow always goes through the memory cells and is essential to keep track of this recurrent path. For this purpose, the gradient is split according to the effect of memory cell on the error $\frac{\partial E(t)}{\partial s_c(t)}$ and the effect the weights have on the memory cell $\frac{\partial s_c(t)}{\partial w_l^m}$. At the same time a RTRL approach is adopted to keep track of the change per memory cell.

Splitting out $\frac{\partial E(t)}{\partial s_c(t)}$

$$\begin{aligned} \frac{\partial E(t)}{\partial s_c(t)} &= \sum_k \frac{\partial E(t)}{\partial o_k(t)} \frac{\partial o_k(t)}{\partial ym_c(t)} \frac{\partial ym_c(t)}{\partial s_c(t)} && \text{(A.15)} \\ &= \sum_k e_k(t) f'(net_{o_k}(t)) g_j^O(t) w_k^c h'(s_c(t)) && \text{(A.16)} \\ &= g_j^O(t) h'(s_c(t)) \sum_k e_k(t) f'(net_{o_k}(t)) w_k^c && \text{(A.17)} \\ & && \text{(A.18)} \end{aligned}$$

and $\frac{\partial s_c(t)}{\partial w_l^m}$:

$$\frac{\partial s_c(t)}{\partial w_l^m} = g_c^F(t) \frac{\partial s_c(t-1)}{\partial w_l^m} + g_c^I(t) \frac{\partial 2 \tanh(net_c(t))}{\partial w_l^m} + 2 \tanh(net_c(t)) \frac{\partial g_c^I(t)}{\partial w_l^m} + s_c(t-1) \frac{\partial g_c^F(t)}{\partial w_l^m} \tag{A.19}$$

Out of the 4 terms in $\frac{\partial s_c(t)}{\partial w_l^m}$, however only the first term is non-zero for all types of units. The other three terms are non-zero each for only exactly one type. Thus the equations for each of the types of units simplify to:

Input gating units:

$$\frac{\partial s_c(t)}{\partial w_l^m} = g_c^F(t) \frac{\partial s_c(t-1)}{\partial w_l^m} + 2 \tanh(net_c(t)) f'(net_{g_c^I}(t)) y_m(t-1) \tag{A.20}$$

Forget gating units:

$$\frac{\partial s_c(t)}{\partial w_l^m} = g_c^F(t) \frac{\partial s_c(t-1)}{\partial w_l^m} + s_c(t-1) f'(net_{g_c^F}(t)) y_m(t-1) \tag{A.21}$$

Memory input units:

$$\frac{\partial s_c(t)}{\partial w_l^m} = g_c^F(t)\frac{\partial s_c(t-1)}{\partial w_l^m} + g_c^I(t)\frac{\partial 2\tanh(net_c(t))}{\partial w_l^m} \tag{A.22}$$

Due to the way the error signal is truncated at the point when it re-enters the memory cells, these equations can be efficiently evaluated at each time step by keeping track of the partial derivatives of a given moment. As the equations are recursive, there is only a need to store them for the current and previous time steps. Thus, the memory requirement only scales with the number of weights in the network and not the number of training time steps, making it possible to calculate the gradients efficiently.

A.5 A comparator network in the LSTM

As part of the simulations for the variables featuring LSTM in chapter 6.3, a comparator network structure was included in the LSTM network. Just like in the shaping experiments, the network was modularised along the unit of a memory module. However, rather than train it through shaping, network weights were set by hand for simplicity and described below.

The comparator network consisted of 3 memory modules, each with 9 cells. As the operation of instantaneous comparison does not require the specialised functionality of memory, and instead is easily achieved in a standard two multi layer feed-forward network, the input and output gates were tied open, while the forget gate was tied down to disable recurrency. The three modules were arranged such that they formed a feed-forward network of depth 2 with 18 cells on the first layer and 9 on the second within the overall network. Each of the cells on the second level provides the output of the comparison between the input vector and one of the stimulus mapping memory modules.

The weights of each of the ith first layer cells had a structure of

$$net_i = \sum_j (w_j x_j^{in} - w_j x_j^{mem_i}) \tag{A.23}$$

Weight values w_j were chosen such that no valid combination summed to the same value, thus only if $\mathbf{x}^{in}$ and $\mathbf{x}^{mem_i}$ are equal is $net_i = 0$. With input and mapping vectors known to be unary encoded this is the case, as long as $w_j \neq w_k$ for $i \neq k$.

The weights of the second set of first layer cells were chosen to be equal and opposite such that $net_i^+ = -net_i^-$

$$cell_i = \sigma(net_i - b) \tag{A.24}$$

Weights were chosen large with respect to the β temperature of the sigmoidal output non-linearity, such that it acted to binarise with a threshold of $net_i > 0$.

Given that the sigmoid for memory cells in the LSTM network ranges between -1 and 1, the two cells $cell_i^+$ and $cell_i^-$ are either -1/+1 or +1/-1 for when the ith mapping module was different from the input vector, and -1/-1 for when they are equal. This results in a binary decision of equality.

Overall, this sub-network provides a remapped input vector upon which the rules can be learnt independent of the actual encoding of the stimuli.

Bibliography

Ucl psychology subject pool. URL `http://uclpsychology.sona-systems.com/`.

P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, ICML '04, pages 1–, New York, NY, USA, 2004. ACM. (page 12)

W. C. Abraham. Metaplasticity: tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, 9(5):387, 2008. (page 80)

W. C. Abraham and M. F. Bear. Metaplasticity: the plasticity of synaptic plasticity. *Trends in Neurosciences*, 19(4):126–130, 1996. (page 80)

J. B. Aimone, W. Deng, and F. H. Gage. Adult neurogenesis: integrating theories and separating functions. *Trends in Cognitive Sciences*, 14(7):325 – 337, 2010. (pages 80 and 147)

R. L. Albin, A. B. Young, and J. B. Penney. The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, 12(10):366 – 375, 1989. (page 38)

G. E. Alexander and M. D. Crutcher. Functional architecture of basal ganglia circuits: neural substrates of parallel processing. *Trends in Neurosciences*, 13(7):266 – 271, 1990. (page 38)

G. E. Alexander, M. R. DeLong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9(1):357–381, 1986. (page 38)

G. I. Allen and N. Tsukahara. Cerebrocerebellar communication systems. *Physiological Reviews*, 54(4):957, 1974. (page 38)

A. Allport, E. Styles, and S. Hsieh. Shifting intentional set: Exploring the dynamic control of tasks. *Attention and Performance XV*, 15:421–452, 1994. (page 84)

G. A. Alvarez and P. Cavanagh. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15(2):106, 2004. (page 117)

J. Anderson. ACT: A simple theory of complex cognition. *American Psychologist*, 51(4):355–365, 1996. (page 10)

J. R. Anderson. Acquisition of cognitive skill. *Psychological Review*, 89(4): 369–406, 1982. (page 119)

J. R. Anderson, D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere, and Y. Qin. An integrated theory of the mind. *Psychological Review*, 111(4): 1036–1060, 2004. (page 10)

J. Annett. Recent developments in hierarchical task analysis. *Contemporary Ergonomics*, pages 263–268, 1996. (page 144)

W. F. Asaad, G. Rainer, and E. K. Miller. Task-specific neural activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 84(1):451, 2000. (page 36)

M. Asadi and M. Huber. Effective control knowledge transfer through learning skill and representation hierarchies. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 2054–2059, 2007. (page 28)

B. B. Averbeck, J.-W. Sohn, and D. Lee. Activity in prefrontal cortex during dynamic selection of action sequences. *Nature Neuroscience*, 2006. (page 25)

A. D. Baddeley. *Working memory*. Oxford University Press, 1986. (pages 34 and 35)

A. D. Baddeley and R. Logie. Working memory: The multiple component model. In A. Miyake and P. Shah, editors, *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press, 1999. (page 117)

A. D. Baddeley, G. J. Hitch, and G. H. Bower. The psychology of learning and motivation. In *Recent advances in learning and motivation*, pages 47–89. Academic Press, 1974. (page 36)

D. Badre and M. D'Esposito. Is the rostro-caudal axis of the frontal lobe hierarchical? *Nature Reviews Neuroscience*, 10(9):659–669, 2009. (page 37)

B. Bakker and J. Schmidhuber. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In F. Groen, N. Amato, A. Bonarini, E. Yoshida, and B. Krse, editors, *Proceedings of the 8th Conference on Intelligent Autonomous Systems*, volume IAS-8, pages 438–445, 2004. (page 81)

B. W. Balleine and A. Dickinson. Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5):407 – 419, 1998. (page 35)

M. T. Banich. The missing link: The role of interhemispheric interaction in attentional processing,. *Brain and Cognition*, 36(2):128 – 157, 1998. (page 20)

M. T. Banich, M. P. Milham, R. Atchley, N. J. Cohen, A. Webb, T. Wszalek, A. F. Kramer, Z. P. Liang, A. Wright, J. Shenker, et al. fMRI studies of Stroop tasks reveal unique roles of anterior and posterior brain systems in attentional selection. *Journal of Cognitive Neuroscience*, 12(6):988–1000, 2000. (page 84)

I. Bar-Gad, G. Morris, and H. Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71(6):439 – 473, 2003. (page 38)

D. M. Barch, T. S. Braver, C. S. Carter, R. A. Poldrack, and T. W. Robbins. CNTRICS final task selection: Executive control. *Schizophrenia Bulletin*, (1):115–135, 2009a. (page 84)

D. M. Barch, C. S. Carter, A. Arnsten, R. W. Buchanan, J. D. Cohen, M. Geyer, M. F. Green, J. H. Krystal, K. Nuechterlein, T. Robbins, et al. Selecting paradigms from cognitive neuroscience for translation into use in clinical trials: proceedings of the third CNTRICS meeting. *Schizophrenia Bulletin*, 35(1):109–114, 2009b. (pages 84 and 145)

A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, 13 (4):341–379, 2003. (pages 78 and 81)

A. G. Barto, R. S. Sutton, and C. W. Anderson. *Neuronlike adaptive elements that can solve difficult learning control problems*, pages 535–549. MIT Press, Cambridge, MA, USA, 1988. (page 30)

J. Baxter. A Bayesian/information theoretic model of learning to learn via multiple task sampling. *Machine Learning*, 28(1):7–39, 1997. (pages 11, 33, and 149)

D. G. Beiser and J. C. Houk. Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology*, 79(6): 3168–3188, 1998. (pages 40 and 42)

Y. Bengio, P. Frasconi, and P. Simard. The problem of learning long-term dependencies in recurrent networks. In *Proceedings of the IEEE International Conference on Neural Networks*, 1993. (pages 40 and 47)

Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *Neural Networks, IEEE Transactions on*, 5 (2):157–166, 1994. (page 47)

Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM New York, NY, USA, 2009. (pages 12 and 30)

K. Berman, J. Ostrem, C. Randolph, J. Gold, T. Goldberg, R. Coppola, R. Carson, P. Herscovitch, and D. Weinberger. Physiological activation of a cortical network during performance of the Wisconsin Card Sorting Test: a positron emission tomography study. *Neuropsychologia*, 33(8): 1027–1046, 1995. (page 85)

J. L. Bermudez. Animal reasoning and proto-logic. *Rational Animals*, pages 127–138, 2006. (page 123)

D. S. Bernstein. Reusing old policies to accelerate learning on new MDPs. Technical report, University of Massachusetts Amherst, 1999. (page 27)

E. Bienenstock, L. Cooper, and P. Munro. Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(1):32, 1982. (page 80)

D. Bindra. A unified account of classical conditioning and operant training. In Editor, editor, *Classical Conditioning II*, pages 453–481. Appleton-Century-Crofts, New York, 1972. (page 23)

D. Bindra. A motivational view of learning, performance, and behavior modification. *Psychological Review*, 81(3):199 – 213, 1974. (page 23)

E. Bleuler. *Dementia Praecox or the group of Schizophrenias*. International Universities Press, 1911. (page 84)

E. Bleuler. *Textbook of Psychiatry*. Arno Press, 1924. (page 84)

J. P. Bolam, J. J. Hanley, P. A. C. Booth, and M. D. Bevan. Synaptic organisation of the basal ganglia. *Journal of Anatomy*, 196(04):527–542, 2000. (page 38)

M. M. Botvinick. Multilevel structure in behaviour and in the brain: a model of Fuster's hierarchy. *Philosophical Transactions of the Royal Society of London*, 2007. (page 141)

M. M. Botvinick, Y. Niv, and A. C. Barto. Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, 113(3):262 – 280, 2009. (page 12)

J. S. Bowers. On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, 116(1):220, 2009. (page 147)

D. H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10(4):433–436, 1997. (page 86)

J. Brandt and N. Butters. Neuropsychological characteristics of Huntington's disease. In L. Grant and K. M. Adams, editors, *Neuropsychological Assessment of Neuropsychiatric Disorders*, pages 312–341. Oxford University Press, 1996. (page 39)

C. Braun, U. Heinz, R. Schweizer, K. Wiech, N. Birbaumer, and H. Topka. Dynamic organization of the somatosensory cortex induced by motor activity. *Brain*, 124(11):2259–2267, 2001. (page 26)

D. A. Braun, C. Mehring, and D. M. Wolpert. Structure learning in action. *Behavioural Brain Research*, 206(2):157 – 165, 2010. (page 149)

T. S. Braver and J. D. Cohen. On the control of control: The role of dopamine in regulating prefrontal function and working memory. In *Control of cognitive processes: Attention and performance XVIII*, pages 713–737, 2000. (page 41)

T. S. Braver, J. D. Cohen, L. E. Nystrom, J. Jonides, E. E. Smith, and D. C. Noll. A parametric study of prefrontal cortex involvement in human working memory. *NeuroImage*, 5:49–62, 1997. (page 83)

T. S. Braver, D. M. Barch, W. M. Kelly, R. L. Buckner, N. J. Cohen, F. Mienzin, A. Z. Snyder, J. M. Ollinger, E. Akbudak, T. E. Conturo, and S. E. Petersen. Direct comparison of prefrontal cortex regions engaged in working and long-term memory tasks. *NeuroImage*, 14:48–59, 2001. (page 83)

K. Breland and M. Breland. The misbehavior of organisms. *American Psychologist*, 16(11):681–684, 1961. (pages 23 and 80)

J. W. Brown and T. S. Braver. Learned predictions of error likelihood in the anterior cingulate cortex. *Science*, 307(5712):1118–1121, 2005. (page 141)

R. G. Brown and C. D. Marsden. Cognitive function in Parkinson's disease: from description to theory. *Trends in Neurosciences*, 13(1):21–29, 1990. (page 42)

N. Brunel and X. J. Wang. Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. *Journal of Computational Neuroscience*, 11(1):63–85, 2001. (page 41)

J. S. Bruner. Going beyond the information given. In *Contemporary Approaches to Cognition: A symposium held at the University of Colorado*, pages 123–152. Harvard University Press, 1957. (page 10)

J. S. Bruner. The act of discovery. *Harvard Educational Review*, 31(1):21–32, 1961. (page 10)

J. Buee, J. M. Deniau, and G. Chevalier. Nigral modulation of cerebello-thalamo-cortical transmission in the ventral medial thalamic nucleus. *Experimental Brain Research*, 65:241–244, 1986. (page 40)

D. Bullock and S. Grossber. Neural dynamics of planned arm movements: Emergent invariants and speed-accuracy properties during trajectory formation. *Psychology Review*, 95(1):49–90, 1988. (page 42)

P. W. Burgess. Assessment of executive function. In *The Handbook of Clinical Neuropsychology*, pages 349–369. Oxford Scholarship Online Monographs, 2010. (page 55)

P. W. Burgess, A. Quayle, and C. D. Frith. Brain regions involved in prospective memory as determined by positron emission tomography. *Neuropsychologia*, 39(6):545 – 555, 2001. (page 36)

P. W. Burgess, I. Dumontheil, and S. J. Gilbert. The gateway hypothesis of rostral prefrontal cortex (area 10) function. *Trends in Cognitive Sciences*, 11(7):290 – 298, 2007a. (pages 37 and 120)

P. W. Burgess, S. J. Gilbert, and I. Dumontheil. Function and localization within rostral prefrontal cortex (area 10). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1481):887–899, 2007b. (page 37)

C. M. Butter. Perseveration in extinction and in discrimination reversal tasks following selective frontal ablations in Macaca mulatta. *Physiology & Behaviour*, 4(2):163–171, 1969. (page 71)

N. Butters, D. Salmon, and W. C. Heindel. Specificity of the memory deficits associated with basal ganglia dysfunction. *Revue Neurologique*, 150 (8-9):580–587, 1994. (page 39)

J. B. Carman, W. M. Cowan, and T. P. S. Powell. The organization of cortico-striate connexions in the rabbit. *Brain*, 86(3):525, 1963. (page 38)

G. Carpenter and S. Grossberg. Adaptive resonance theory: Stable self-organization of neural recognition codes in response to arbitrary lists of input patterns, 1986. (page 26)

G. A. Carpenter and S. Grossberg. The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88, 1988. (pages 40, 60, 79, and 146)

R. Case and C. Bereiter. From behaviourism to cognitive behaviourism to cognitive development: Steps in the evolution of instructional design. *Instructional Science*, 13:141–158, 1984. (page 25)

M. V. Chafee and P. S. Goldman-Rakic. Matching patterns of activity in primate prefrontal area 8a and parietal area 7ip neurons during a spatial working memory task. *Journal of Neurophysiology*, 79(6):2919, 1998. (page 36)

S. Charron and E. Koechlin. Divided representation of concurrent goals in the human frontal lobes. *Science*, 328(5976):360–363, 2010. (page 150)

N. Chater and C. Heyes. Animal concepts: Content and discontent. *Mind & Language*, 9(3):209–246, 1994. (page 123)

G. Chevalier and J. M. Deniau. Disinhibition as a basic process in the expression of striatal functions. *Trends in Neuroscience*, 13(7):277–280, 1990. (pages 40 and 42)

G. Chevalier, S. Vacher, J. M. Deniau, and M. Desban. Disinhibition as a basic process in the expression of striatal functions. I. The striato-nigral influence on tecto-spinal/tecto-diencephalic neurons. *Brain Research*, 334(2): 215 – 226, 1985. (page 40)

E. J. Coan, A. J. Irving, and G. L. Collingridge. Low-frequency activation of the NMDA receptor system can prevent the induction of LTP. *Neuroscience letters*, 105(1-2):205–210, 1989. (page 80)

J. D. Cohen and D. Servan-Schreiber. Context, cortex, and dopamine: a connectionist approach to behavior and biology in schizophrenia. *Psychological Review*, 99(1):45–77, 1992. (page 83)

J. D. Cohen, T. S. Braver, and R. C. O'Reilly. A computational approach to prefrontal cortex, cognitive control and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society of London*, 351:1515–1527, 1996. (page 40)

J. D. Cohen, W. M. Perlstein, T. S. Braver, L. E. Nystrom, D. C. Noll, J. Jonides, and E. E. Smith. Temporal dynamics of brain activation during a working memory task. *Nature*, 386(6625):604 – 608, 1997. (page 83)

J. D. Cohen, D. M. Barch, C. Carter, and D. Servan-Schreiber. Context-processing deficits in schizophrenia: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology*, 108:120–133, 1999. (page 83)

J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automated processing: A parallel distributed processing account of the stroop effect. *Psychological Review*, 97(3):332–361, 2000. (pages 37 and 141)

J. D. Cohen, T. S. Braver, and J. W. Brown. Computational perspectives on dopamine function in prefrontal cortex. *Current Opinion in Neurobiology*, 12:223–229, 2002. (page 141)

M. X. Cohen and M. J. Frank. Neurocomputational models of basal ganglia function in learning, memory and choice. *Behavioural Brain Research*, 199(1): 141 – 156, 2009. (page 141)

S. M. Courtney, L. Petit, J. M. Maisog, L. G. Ungerleider, and J. V. Haxby. An area specialized for spatial working memory in human frontal cortex. *Science*, 279(5355):1347, 1998. (page 36)

N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(01):87–114, 2001. (pages 35, 117, and 147)

N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal system for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005. (pages 119, 120, and 140)

P. Dayan. Images, frames, and connectionist hierarchies. *Neural Computation*, 18(10):2293–2319, 2006. (page 80)

P. Dayan. Bilinearity, rules, and prefrontal cortex. *Frontiers in Computational Neuroscience*, 1, 2007. (pages 40, 44, 52, 53, 54, 57, 82, 126, 130, 131, 145, and 148)

P. Dayan and Y. Niv. Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196, 2008. (page 43)

P. Dayan and A. J. Yu. Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems*, 17(4):335–350, 2006. (page 74)

E. de Villers-Sidani, K. L. Simpson, Y. F. Lu, R. C. S. Lin, and M. M. Merzenich. Manipulating critical period closure across different sectors of the primary auditory cortex. *Nature Neuroscience*, 11(8):957–965, 2008. (page 149)

S. Dehaene and J.-P. Changeux. A hierarchical neuronal network for planning behavior. *Proceedings of the National Academy of Sciences*, 94:13293–13298, 1997. (page 141)

M. R. Delgado, L. E. Nystrom, C. Fissell, D. C. Noll, and J. A. Fiez. Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, 84(6):3072, 2000. (page 40)

G. J. Demakis. A meta-analytic review of the sensitivity of the Wisconsin Card Sorting Test to frontal and lateralized frontal brain damage. *Neuropsychology*, 17(2):255–264, 2003. (page 85)

w. Deng, J. B. Aimone, and F. H. Gage. New neurons and new memories: how does adult hippocampal neurogenesis affect learning and memory? *Nature Reviews Neuroscience*, 11(5):339–350, 2010. (page 147)

M. D'Esposito. Chapter 11 working memory. In M. J. Aminoff, F. Boller, D. F. Swaab, G. Goldenberg, and B. L. Miller, editors, *Neuropsychology and Behavioral Neurology*, volume 88 of *Handbook of Clinical Neurology*, pages 237 – 247. Elsevier, 2008. (page 34)

M. D'Esposito, B. Postle, and B. Rypma. Prefrontal cortical contributions to working memory: Evidence from event-related fMRI studies. *Experimental Brain Research*, 133(1):3–11, 2000. (page 35)

B. D. Devan and N. M. White. Parallel information processing in the dorsal striatum: Relation to hippocampal function. *Journal of Neuroscience*, 19(7): 2789–2798, 1999. (page 119)

A. Dickinson. Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 308(1135):67–78, 1985. (page 119)

A. Dickinson and B. Balleine. The role of learning in the operation of motivational systems. In C. R. Gallistel, editor, *Stevens Handbook of Experimental Psychology*, volume 3, page 497533. Wiley Online Library, 2002. (page 39)

S. Diekelmann and J. Born. The memory function of sleep. *Nature Reviews Neuroscience*, 2010. (page 118)

B. B. Doll, W. J. Jacobs, A. G. Sanfey, and M. J. Frank. Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, 1299:74 – 94, 2009. (page 120)

B. B. Doll, K. E. Hutchison, and M. J. Frank. Dopaminergic genes predict individual differences in susceptibility to confirmation bias. *The Journal of Neuroscience*, 31(16):6188, 2011. (page 13)

P. F. Dominey and M. A. Arbib. Cortico-subcortical model of generation of spatially accurate sequential saccades. *Cereberal Cortex*, 2(2):152 – 175, 1992. (page 42)

M. Dorigo and M. Colombetti. Robot shaping: Developing situated agents through learning. Technical Report TR-92-040, International Computer Science Institute, April 1993. (pages 31 and 32)

M. Dorigo and M. Colombetti. *Robot Shaping: An Experiment in Behavior Engineering*. MIT Press/Bradford Books, 1998. (pages 12, 28, and 31)

J. C. Dreher, E. Guigon, and Y. Burnod. A model of prefrontal cortex dopaminergic modulation during the delayed alternation task. *Journal of Cognitive Neuroscience*, 14(6):853–865, 2002. (page 41)

B. Dubois and B. Pillon. Cognitive deficits in Parkinson's disease. *Journal of Neurology*, 244(1):2–8, 1996. (page 39)

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2nd edition, 2000. (pages 11 and 30)

I. Dumontheil, R. Thompson, and J. Duncan. Assembly and use of new task rules in fronto-parietal cortex. *Journal of Cognitive Neuroscience*, 23(1): 168–182, 2011. (page 121)

J. Duncan. Attention, intelligence, and the frontal lobes. In M. S. Gazzaniga, editor, *The New Cognitive Neurosciences*, pages 721–733. The MIT Press, 1995. (page 81)

J. Duncan, H. Emslie, P. Williams, R. Johnson, and C. Freer. Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive Psychology*, 30(3):257–303, 1996. (page 121)

J. Duncan, A. Parr, A. Woolgar, R. Thompson, P. Bright, S. Cox, S. Bishop, and I. Nimmo-Smith. Goal neglect and Spearman's g: Competing parts of a complex task. *Journal of Experimental Psychology-General*, 137(1):131–148, 2008. (pages 53 and 121)

D. Durstewitz, J. K. Seamans, and T. J. Sejnowski. Dopamine-mediated stabilization of delay-periods activity in a network model of prefrontal cortex. *The American Physiological Society*, pages 1733–1750, 2000a. (page 41)

D. Durstewitz, J. K. Seamans, and T. J. Sejnowski. Neurocomputational models of working memory. *Nature Neuroscience*, 3:1184–1191, 2000b. (page 35)

M. J. Eacott and D. Gaffan. Inferotemporal-frontal disconnection: The uncinate fascicle and visual associative learning in monkeys. *European Journal of Neuroscience*, 4(12):1320–1332, 1992. (page 37)

H. Ebbinghaus. *Memory: A Contribution to Experimental Psychology*. Teachers College, Columbia University, 1913. (page 10)

D. Eck and J. Schmidhuber. Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In *Neural Networks for Signal Processing XII, Proc. 2002 IEEE Workshop*, pages 747–756, 2002. (page 47)

D. A. Eckerman, R. D. Hienz, S. Stern, and V. Kowlowitz. Shaping the location of a pigeon's peck: effect of rate and size of shaping steps. *Journal of the Experimental Analysis of Behavior*, 33(3):299–310, 1980. (page 11)

J. L. Elman. Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences*, 9(3):111–117, 2005. (page 11)

J. L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993. (pages 31, 32, 59, 81, and 149)

T. Erez and W. Smart. What does shaping mean for computational reinforcement learning? In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on*, pages 215 –219, 2008. (pages 19 and 22)

N. Eshel, J. Luka, A. Lenartowicz, L. E. Nystrom, and J. D. Cohen. Transiently disrupting right prefrontal cortex interferes with updating of working memory. Pre-print manuscript from author, 2009. (pages 84 and 119)

J. Feldman. An algebra of human concept learning. *Journal of Mathematical Psychology*, 50(4):339 – 368, 2006. (page 120)

K. Ferguson and S. Mahadevan. Proto-transfer learning in markov decision processes using spectral methods. In *ICML-06 Workshop on Structural Knowledge Transfer for Machine Learning*, 2006. (page 33)

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3 – 71, 1988. (page 143)

C. A. Fox and J. A. Rafols. The striatal efferents in the globus pallidus and in the substantia nigra. In M. D. Yahr, editor, *The Basal Ganglia*, pages 37 – 55. Raven Press, 1976. (page 38)

D. Frank, M. J. Badre. Mechanisms of hierarchical reinforcement learning in corticostriatal circuits 1: Computational analysis. *Cerebral Cortex*, 2011. (page 141)

M. J. Frank. Hold your horses: A dynamic computational role for the subthalamic nucleus in decision making. *Neural Networks*, 19(8):1120–1136, 2006. (page 141)

M. J. Frank and K. Hutchison. Genetic contributions to avoidance-based decisions: striatal d2 receptor polymorphisms. *Neuroscience*, 164(1):131 – 140, 2009. (page 13)

M. J. Frank, B. Loughry, and R. C. O'Reilly. Interactions between frontal cortex and basal ganglia in working memory: A computational model. Technical report, Department of Psychology, University of Colorado, 2001a. (pages 36, 41, 44, 56, 57, 63, 82, 145, and 147)

M. J. Frank, B. Loughry, and R. C. O'Reilly. Interactions between frontal cortex and basal ganglia in working memory: a computational model. *Cognitive, Affective, Behavioral Neuroscience*, 1(21):137–160, 2001b. (page 119)

M. J. Frank, L. C. Seeberger, and R. C. O'Reilly. By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, 306:1940, 2004. (page 141)

R. M. French. Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, 4(3):365–377, 1992. (page 27)

R. M. French. Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 335–340, 1994. (pages 27 and 142)

R. M. French. Catastrophic forgetting in connectionist networks: Causes, consequences and solutions. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. (pages 12, 60, and 142)

S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex. *Journal of Neurophysiology*, 61(2):331, 1989. (page 36)

S. Funahashi, C. J. Bruce, and P. S. Goldman-Rakic. Dorsolateral prefrontal lesions and oculomotor delayed-response performance: Evidence for mnemonic "scotomas". *J. Neuroscience*, 13(4):1479–1497, 1993. (page 36)

S. Fusi, P. J. Drew, and L. F. Abbott. Cascade models of synaptically stored memories. *Neuron*, 45(4):599–611, 2005. (page 80)

J. M. Fuster. *The Prefrontal Cortex*. Academic Press, 2008. (page 35)

J. M. Fuster. *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe*. Lippincott-Raven, 3rd edition, 1989. (page 35)

J. M. Fuster and G. E. Alexander. Neuron activity related to short-term memory. *Science*, 173(3997):652, 1971. (page 36)

J. M. Fuster and J. P. Jervey. Inferotemporal neurons distinguish and retain behaviorally relevant features of visual stimuli. *Science*, 212(4497):952–955, 1981. (page 36)

R. M. Gagné. Military training and principles of learning. *American Psychologist*, 17(2):83 – 91, 1962. (pages 24, 28, and 151)

R. M. Gagné. *The conditions of learning*. Holt, Rinehart and Winston, 1970. (pages 24 and 144)

F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000. (pages 43, 46, 47, 50, and 164)

F. A. Gers, D. Eck, and J. Schmidhuber. Applying LSTM to time series predictable through time-window approaches. *Artificial Neural NetworksICANN 2001*, pages 669–676, 2001. (page 47)

F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research*, 3:115–143, 2003. (pages 43 and 47)

A. Gevins and B. Cutillo. Spatiotemporal dynamics of component processes in human working memory. *Electroencephalography and Clinical Neurophysiology*, 87(3):128–143, 1993. (page 83)

S. J. Gilbert and P. W. Burgess. Executive function. *Current Biology*, 18 (3):R110–R114, 2008. (page 36)

J. Gläscher, N. D. Daw, P. Dayan, and J. O'Doherty. States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, 66(4):585–595, 2010. (page 120)

P. S. Goldman and H. E. Rosvold. The effects of selective caudate lesions in infant and juvenile rhesus monkeys. *Brain Research*, 43(1):53 – 66, 1972. (page 39)

P. S. Goldman-Rakic. Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handbook of Psychology - The Nervous System*, 5:373 – 417, 1987. (pages 35 and 37)

P. S. Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477–485, 1995. (page 36)

P. S. Goldman-Rakic. Regional and cellular fractionation of working memory. *Proceedings of the National Academy of Sciences of the United States of America*, 93(24):13473–13480, 1996. (page 36)

R. L. Gomez. Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, 33(2):154, 1997. (page 85)

N. D. Goodman, J. B. Tenenbaum, J. Feldman, and T. L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1): 108–154, 2008. (page 120)

D. Gopher, L. Armony, and Y. Greenshpan. Switching tasks and attention policies. *Journal of Experimental Psychology General*, 129(3):308–339, 2000. (page 84)

U. Goswami. Neuroscience and education: from research to practice? *Nature Reviews Neuroscience*, 7:406–413, 2006. (page 151)

U. Goswami and D. Szucs. Educational neuroscience: Developmental mechanisms: Towards a conceptual framework. *NeuroImage*, In Press, Corrected Proof:–, 2010. (page 151)

A. M. Gotham, R. G. Brown, and C. D. Marsden. 'Frontal' cognitive function in patients with Parkinson's disease 'on' and 'off' Levodopa. *Brain*, 111(2):299–321, 1988. (page 39)

E. Gould, A. Beylin, P. Tanapat, A. Reeves, and T. J. Shors. Learning enhances adult neurogenesis in the hippocampal formation. *Nature Neuroscience*, 2(3):260–265, 1999. (page 147)

D. A. Grant and E. A. Berger. A behavioural analysis of degree of reinforcement and ease of shifting to new responses in a weigl type card sorting problem. *Journal of Experimental Psychology*, 38:404 – 411, 1948. (page 85)

A. Graves and J. Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. *Advances in Neural Information Processing Systems*, 21, 2009. (page 47)

A. Graves, D. Eck, N. Beringer, and J. Schmidhuber. Biologically plausible speech recognition with LSTM neural nets. In A. J. Ijspeert, M. Murata, and N. Wakamiya, editors, *Biologically Inspired Approaches to Advanced Information Technology*, volume 3141 of *Lecture Notes in Computer Science*, pages 127–136. Springer Berlin / Heidelberg, 2004. (page 47)

A. M. Graybiel. The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory*, 70(1):119–136, 1998. (page 39)

S. Grossberg. How does a brain build a cognitive code? *Psychological Review*, 87(1):1–51, 1980. (pages 26, 40, 60, 65, and 79)

A. J. Gruber, P. Dayan, B. S. Gutkin, and S. A. Solla. Dopamine modulation in the basal ganglia locks the gate to working memory. *Journal of Computational Neuroscience*, 20(2):153–166, 2006. (page 43)

V. Gullapalli. *Reinforcement learning and its application to control*. PhD thesis, University of Massachusetts, Amherst, MA, USA, 1992. (pages 28 and 31)

V. Gullapalli. Reinforcement learning of complex behavior through shaping. In J. W. Donahoe and V. P. Dorsel, editors, *Neural-Network Models of Cognition - Biobehavioral Foundations*, volume 121 of *Advances in Psychology*, pages 302 – 314. North-Holland, 1997. (page 31)

S. N. Haber. The primate basal ganglia: Parallel and integrative networks. *Journal of Chemical Neuroanatomy*, 26(4):317 – 330, 2003. (page 38)

R. F. Hadley. Systematicity in connectionist language learning. *Mind Language*, 9(3):247–272, 1994. (page 143)

R. Hadsell, A. Erkan, P. Sermanet, M. Scoffier, U. Muller, and Y. LeCun. Deep belief net learning in a long-range vision system for autonomous offroad driving. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 628 –633, sept. 2008. (page 143)

D. L. Harrington and K. Y. Haaland. Sequencing in Parkinson's disease: abnormalities in programming and controlling movement. *Brain*, 114(1):99, 1991. (page 42)

C. L. Harris. *Parallel distributed processing models and metaphors for language and development*. PhD thesis, University of California at San Diego, La Jolla, CA, USA, 1992. (page 31)

M. Haruno, D. M. Wolpert, and M. Kawato. MOSAIC model for sensorimotor learning and control. *Neural Computation*, 13(10):2201–2220, 2001. (pages 79 and 146)

T. E. Hazy, W. Pauli, S. Herd, and R. C. OReilly. Neural mechanisms of executive function: A biological and computational framework. (page 20)

T. E. Hazy, M. J. Frank, and R. C. O'Reilly. Banishing the homunculus: Making working memory work. *Neuroscience*, 2005. (pages 36 and 44)

T. E. Hazy, M. J. Frank, and R. C. O'Reilly. Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society B. Biological Sciences*, 362(1485):1601–1613, 2007. (pages 44, 57, 82, and 145)

D. O. Hebb. *The Organization of Behavior*. Wiley, New York, 1949. (pages 18 and 35)

S. A. Herd, M. T. Banich, and R. C. O'reilly. Neural mechanisms of cognitive control: An integrative model of stroop task performance and fmri data. *Journal of Cognitive Neuroscience*, 18(1):22–32, 2006. (page 141)

G. E. Hinton. Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, 1:143 – 150, 1989. (page 45)

G. E. Hinton. Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, 46(1-2):47–75, 1990. (page 124)

G. E. Hinton and Z. Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B*, 352(1358):1177–1190, 1997. (page 81)

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, chapter Distributed representations, pages 77–109. MIT Press, 1986. (pages 11 and 26)

G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006. (pages 81 and 143)

S. Hochreiter. Untersuchungen zu dynamischen Neuronalen Netzen. Master's thesis, Technische Universität München, 1991. (page 47)

S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. (pages 40, 43, 46, 60, and 166)

J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences of the United States of America*, 81(10):3088, 1984. (page 35)

E. Hoshi, K. Shima, and J. Tanji. Task-dependent selectivity of movement-related neuronal activity in the primate prefrontal cortex. *Journal of Neurophysiology*, 80(6):3392, 1998. (pages 36 and 149)

J. C. Houk and S. P. Wise. Feature article: Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex*, 5(2):95–110, 1995. (page 42)

J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. In J. C. Houk, J. Davis, and D. Beiser, editors, *Models of information processing in the basal ganglia*, pages 249–270. MIT press, 1995a. (page 42)

J. C. Houk, J. L. Davis, and D. G. Beiser. *Models of information processing in the basal ganglia*. The MIT press, 1995b. (page 38)

A. Hyafil, C. Summerfield, and E. Koechlin. Two mechanisms for task switching in the prefrontal cortex. *Journal of Neuroscience*, 29(16):5135–5142, 2009. (page 84)

S. D. Iversen and M. Mishkin. Perseverative interference in monkeys following selective lesions of the inferior prefrontal convexity. *Experimental Brain Research*, 11(4):376–386, 1970. (page 71)

R. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991a. (pages 79 and 146)

R. A. Jacobs, M. I. Jordan, and A. G. Barto. Task decomposition through competition in a modular connectionist architecture: The what and where vision tasks. *Cognitive Science*, 15:219–250, 1991b. (page 29)

W. James. *The principles of psychology (Vols. 1 & 2)*. Holt, New York, 1890. (page 35)

A. T. Jersild. Mental set and shift. *Archives of Psychology. Vol*, 14(89):81, 1927. (page 84)

A. P. Jha and G. McCarthy. The influence of memory load upon delay-interval activity in a working-memory task: An event-related functional MRI study. *Journal of Cognitive Neuroscience*, 12(2):90–105, 2000. (page 36)

D. Joel, Y. Niv, and E. Ruppin. Actor - critic models of the basal ganglia: new anatomical and computational perspectives. *Neural Networks*, 15:535–547, 2004. (pages 40 and 52)

B. Jones and M. Mishkin. Limbic lesions and the problem of stimulus–reinforcement associations. *Experimental Neurology*, 36(2):362–377, 1972. (page 71)

M. Jones and R. Sugden. Positive confirmation bias in the acquisition of information. *Theory and Decision*, 50:59–99, 2001. (page 145)

S. Káli and P. Dayan. Off-line replay maintains declaritive memory in a model. *Nature Neuroscience*, 7(3):286–294, 2004. (page 146)

S. W. Keele, R. Ivry, U. Mayr, E. Hazeltine, and H. Heuer. The cognitive and neural architecture of sequence representation. *Psychological Review*, 110 (2):316–339, 2003. (page 83)

C. Kemp and J. B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692, 2008. (page 149)

J. M. Kemp and T. P. S. Powell. The cortico-striate projection in the monkey. *Brain*, 93(3):525, 1970. (page 38)

A. Kincaid, T. Zheng, and C. Wilson. Connectivity and convergence of single corticostriatal axons. *Journal of Neuroscience*, 18(12):4722, 1998. (page 42)

A. Kinder. The knowledge acquired during artificial grammar learning: Testing the predictions of two connectionist models. *Psychological Research*, 63:95–105, 2000. (page 141)

B. J. Knowlton, L. R. Squire, and M. A. Gluck. Probabilistic classification learning in amnesia. *Learning & Memory*, 1(2):106, 1994. (page 123)

B. J. Knowlton, J. A. Mangels, and L. R. Squire. A neostriatal habit learning system in humans. *Science*, 273(5280):1399, 1996. (page 39)

B. Knutson, A. Westdorp, E. Kaiser, and D. Hommer. FMRI visualization of brain activity during a monetary incentive delay task. *NeuroImage*, 12(1): 20 – 27, 2000. (page 40)

E. Koechlin and A. Hyafil. Anterior prefrontal function and the limits of human decision-making. *Science*, 318(5850):594–598, 2007. (page 37)

E. Koechlin and C. Summerfield. An information theoretical approach to prefrontal executive function. *Trends in Cognitive Sciences*, 11(6):229–235, 2007. (pages 119, 141, and 149)

E. Koechlin, G. Basso, P. Pietrini, S. Panzer, and J. Grafman. The role of the anterior prefrontal cortex in human cognition. *Nature*, 399(6732): 148–151, 1999. (page 150)

E. Koechlin, C. Ody, and F. Kouneiher. The architecture of cognitive control in the human prefrontal cortex. *Science*, 302(5648):1181–1185, 2003. (pages 37, 119, and 149)

G. Konidaris and A. Barto. Autonomous shaping: Knowledge transfer in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pages 489–496. ACM Press New York, NY, USA, 2006. (page 33)

C. A. Kortge. Episodic memory in connectionist networks. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, volume 764, page 771, 1990. (pages 26 and 142)

F. Kouneiher, S. Charron, and E. Koechlin. Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, 12(7):939–945, 2009. (page 118)

E. Kraepelin. *Dementia Praecox and Paraphrenia*. E & S Livingston, 1913. (page 84)

R. M. Kretchmar, T. Feil, and R. Bansal. Improved automatic discovery of subgoals for options in hierarchical reinforcement learning. *Journal of Computer Science and Technology*, 3(2), 2003. (page 13)

M. F. Kritzer and P. S. Goldman-Rakic. Intrinsic circuit organization of the major layers and sublayers of the dorsolateral prefrontal cortex in the rhesus monkey. *The Journal of Comparative Neurology*, 359(1):131–143, 1995. (page 36)

K. A. Krueger and P. Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380 – 394, 2009. (page 15)

D. Kumaran, J. J. Summerfield, D. Hassabis, and E. A. Maguire. Tracking the emergence of conceptual knowledge during human decision making. *Neuron*, 63(6):889–901, 2009. (pages 123 and 132)

K. J. Lang, A. H. Waibel, and G. E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23–43, 1990. (page 47)

H. Lange, G. Thorner, and A. Hopf. Morphometric-statistical structure analysis of human striatum, pallidum and nucleus su-thalamicus. III. Nucleus subthalamicus. *Journal für Hirnforschung*, 17(1):31, 1976. (page 38)

M. Laubach. Who's on first? What's on second? The time course of learning in corticostriatal systems. *Trends in Neurosciences*, 28(10):509–511, 2005. (page 140)

A. Laud and G. DeJong. Reinforcement learning and shaping: Encouraging intended behaviors. In *Proceedings of the Nineteenth International Conference on Machine Learning*, ICML '02, pages 355–362, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc. (page 31)

J. R. Law, M. A. Flanery, S. Wirth, M. Yanike, A. C. Smith, L. M. Frank, W. A. Suzuki, E. N. Brown, and C. E. L. Stark. Functional magnetic resonance imaging activity during the gradual acquisition and expression of paired-associate memory. *Journal of Neuroscience*, 25(24):5720, 2005. (page 93)

C. Lebiere and J. R. Anderson. A connectionist implementation of the act-r production system. 1993. (page 11)

I. P. Levin, S. L. Schneider, and G. J. Gaeth. All frames are not created equal: A typology and critical analysis of framing effects,. *Organizational Behavior and Human Decision Processes*, 76(2):149 – 188, 1998. (page 144)

J. B. Levitt, D. A. Lewis, T. Yoshioka, and J. S. Lund. Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 and 46). *Journal of Comparative Neurology*, 338(3):360–376, 1993. (pages 36 and 45)

J. B. Levitt, J. S. Lund, and T. Yoshioka. Anatomical substrates for early stages in cortical processing of visual information in the macaque monkey. *Behavioural Brain Research*, 76(1-2):5–19, 1996. (page 45)

G. D. Logan and C. Bundesen. Clever homunculus: Is there an endogenous act of control in the explicit task-cuing procedure? *Journal of Experimental Psychology: Human Perception and Performance*, 29(3):575–599, 2003. (page 84)

M. Luciana, R. A. Depue, P. Arbisi, and A. Leon. Facilitation of working memory in humans by a D2 dopamine receptor agonist. *Journal of Cognitive Neuroscience*, 4(1):58–68, 1992. (page 41)

S. J. Luck and E. K. Vogel. The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279–280, 1997. (page 35)

C. M. MacLeod. Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, 109(2):163–203, 1991. (pages 37 and 84)

C. M. MacLeod and K. Dunbar. Training and Stroop-like interference: Evidence for a continuum of automaticity. *J. Experimental Psychology: Learning, Memory and Cognition*, 14(1):126–135, 1988. (page 84)

S. F. Maier. Learned helplessness and animal models of depression. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 8(3):435 – 446, 1984. (page 145)

S. F. Maier and L. R. Watkins. Stressor controllability and learned helplessness: The roles of the dorsal raphe nucleus, serotonin, and corticotropin-releasing factor. *Neuroscience Biobehavioral Reviews*, 29(4-5):829 – 841, 2005. (page 145)

E. Marder and J. M. Goaillard. Variability, compensation and homeostasis in neuron and network function. *Nature Reviews Neuroscience*, 7(7):563–574, 2006. (page 140)

D. Marr. *Vision: A computational approach*. Freeman & Co., San Francisco, 1982. (page 13)

G. Martin, J. Pear, and M. Garry. *Behavior Modification: What it is and how to do it*. Prentice Hall, Upper Saddle River, NJ, 1999. (page 22)

M. Martone, N. Butters, M. Payne, J. T. Becker, and D. S. Sax. Dissociations between skill learning and verbal recognition in amnesia and dementia. *Archives of Neurology*, 41(9):965, 1984. (page 39)

U. Mayr and S. Keele. Changing internal constraints on action: The role of backward inhibition. *Journal of Experimental Psychology: General*, 129(1): 4–26, 2000. (page 84)

J. E. Mazur. *Learning and Behavior*. Prentice Hall, 2005. (page 12)

J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly. 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(2):419–457, 1995. (pages 27, 80, and 142)

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: the sequential learning problem. In G. H. Bower, editor, *The Psychology of Learning and Motivation*, volume 24, pages 169–164. Accademic Press, 1989. (pages 12, 26, 60, and 142)

S. M. McClure, G. S. Berns, and P. R. Montague. Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, 38(2):339 – 346, 2003. (page 40)

S. M. McClure, M. K. York, and P. R. Montague. The neural substrates of reward processing in humans: The modern role of fMRI. *The Neuroscientist*, 10(3):260–268, 2004. (page 40)

R. J. McDonald and N. M. White. A triple dissociation of memory systems: Hippocampus, amygdala, and dorsal striatum. *Behavioral Neuroscience*, 107 (1):3 – 22, 1993. (page 39)

A. McGovern and A. G. Barto. Automatic discovery of subgoals in reinforcement learning using diverse density. In *Proceedings of 18th International Conference on Machine Learning*, pages 361–368, 2001. (pages 13 and 81)

P. McLeod, K. Plunkett, and E. T. Rolls. *Introduction to connectionist modelling of cognitive processes*. Oxford University Press, 1998. (page 11)

K. McRae and P. A. Hetherington. Catastrophic interference is eliminated in pretrained networks. In *Proceedings of the 15h Annual Conference of the Cognitive Science Society*, pages 723–728, 1993. (page 27)

N. Meiran, Z. Chorev, and A. Sapir. Component processes in task switching. *Cognitive Psychology*, 41(3):211–253, 2000. (page 84)

M. Midgley, S. E. G. Lea, and R. M. Kirby. Algorithmic shaping and misbehavior in the acquisition of token deposit by rats. *Journal of the Experimental Analysis of Behavior*, 52(1):27, 1989. (pages 23, 25, and 151)

J. A. Mikels and P. A. Reuter-Lorenz. Neural gate keeping: the role of interhemispheric interactions in resource allocation and selective filtering. *Neuropsychology*, 18(2):328–339, 2004. (page 20)

E. K. Miller and J. D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review Neurosciance*, 24:167–202, 2001. (pages 35 and 37)

E. K. Miller, L. Li, and R. Desimone. A neural mechanism for working and recognition memory in inferior temporal cortex. *Science*, 254(5036):1377, 1991. (page 124)

E. K. Miller, C. A. Erickson, and R. Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16):5154, 1996. (pages 36 and 124)

G. A. Miller. The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956. (pages 35 and 147)

G. A. Miller, E. Galanter, and K. H. Pribram. *Plans and the structure of behavior*. Holt, Rinehart and Winston New York, 1960. (page 35)

B. Milner. Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, 9(1):90, 1963. (pages 36, 85, and 123)

J. W. Mink. The basal ganglia: focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50(4):381–425, 1996. (page 42)

J. W. Mink and W. T. Thach. Basal ganglia intrinsic circuits and their role in behavior. *Current Opinion in Neurobiology*, 3(6):950 – 957, 1993. (page 38)

A. Mohamed, D. Yu, and L. Deng. Investigation of full-sequence training of deep belief networks for speech recognition. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010. (page 143)

S. Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, 2003. (page 84)

S. L. Moody, S. P. Wise, G. di Pellegrino, and D. Zipser. A model that accounts for activity in primate frontal cortex during a delayed matching-to-sample task. *Journal of Neuroscience*, 18(1):399, 1998. (pages 124 and 139)

J. R. Movellan. Contrastive Hebbian learning in the continuous Hopfield model. In *Proceedings of the 1989 Connectionist Models Summer School*, pages 10 – 17, 1990. (page 45)

M. C. Mozer. Induction of multiscale temporal structure. In *NIPS*, volume 4, pages 275 – 282, 1992. (page 40)

G. L. Murphy. *The big book of concepts*. The MIT Press, 2004. (page 122)

J. Murre. The effects of pattern presentation on interference in backpropagation networks. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society*, pages 54–59, 1992. (pages 27 and 142)

Y. Nagahama, T. Okada, Y. Katsumi, T. Hayashi, H. Yamauchi, C. Oyanagi, J. Konishi, H. Fukuyama, and H. Shibasaki. Dissociable mechanisms of attentional control within the human prefrontal cortex. *Cerebral Cortex*, 11(1):85, 2001. (page 85)

V. Nair and G. Hinton. 3d object recognition with deep belief nets. In *Advances in Neural Information Processing Systems*, volume 22, pages 1339–1347, 2009. (page 143)

E. J. Neafsey, C. D. Hull, and N. A. Buchwald. Preparation for movement in the cat. II. unit activity in the basal ganglia and thalamus. *Electroencephalography and Clinical Neurophysiology*, 44(5):714 – 723, 1978. (page 42)

E. L. Newport. Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language. *Language Sciences*, 10(1): 147–172, 1988. (page 31)

E. L. Newport. Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28, 1990. (page 31)

R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175, 1998. (page 145)

M. J. Nissen and P. Bullemer. Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1):1 – 32, 1987. (page 83)

Y. Niv, M. O Duff, and P. Dayan. Dopamine, uncertainty and TD learning. *Behavioral and Brain Functions*, 1(6), 2005. (page 39)

K. A. Norman and R. C. O'Reilly. Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, 110(4):611–646, 2003. (pages 27 and 80)

K. A. Norman, E. L. Newman, and A. J. Perotte. Methods for reducing interference in the complementary learning systems model: Oscillating inhibition and autonomous memory rehearsal. *Neural Networks*, 18(9):1212–1228, 2005. (page 27)

K. H. Nuechterlein and M. E. Dawson. Information processing and attentional functioning in the developmental course of schizophrenia disorders. *Schizophrenia Bulletin*, 10(2):160–203, 1984. (page 83)

J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan. Temporal difference models and reward-related learning in the human brain. *Neuron*, 38(2):329 – 337, 2003. (page 40)

E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267 – 273, 1982. (page 45)

R. C. O'Reilly. Generalization in interactive networks: The benefits of inhibitory competition and hebbian learning. *Neural Computation*, 13:1199 – 1242, 2001. (page 44)

R. C. O'Reilly. Biologically based computational models of higher-level cognition. *Science*, 314:91–94, 2006. (page 141)

R. C. O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2:455–462, 1998. (pages 44 and 147)

R. C. O'Reilly and M. J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2005. (pages 14, 40, 43, 45, 50, 57, 58, 62, 63, 82, 115, 141, and 145)

R. C. O'Reilly and Y. Munakata. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* MIT Press, 2000. (pages 40 and 44)

R. C. O'Reilly and K. A. Norman. Hippocampal and neocortical contributions to memory: Advances in the complimentary learning systems framework. *Trends in Cognitive Science*, 12:505 – 510, 2002. (pages 27 and 80)

R. C. O'Reilly and J. W. Rudy. Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, 108:311–345, 2001. (pages 27 and 80)

R. C. O'Reilly, M. Mozer, Y. Munakata, and A. Miyake. Discrete representations in working memory: a hypothesis and computational investigations. In *The Second International Conference on Cognitive Science*, 1999. (page 40)

R. C. O'Reilly, D. C. Noelle, T. S. Braver, and J. D. Cohen. Prefrontal cortex and dynamic categorisation task: representational organisation and neuromodulatory control. *Cerebral Cortex*, 12:246–257, 2002. (page 141)

R. C. O'Reilly, M. J. Frank, T. E. Hazy, and B. Watz. PVLV: the primary value and learned value Pavlovian learning algorithm. *Behavioral Neuroscience*, 121(1):31–49, 2007. (pages 46 and 50)

R. C. O'Reilly, S. A. Herd, and W. M. Pauli. Computational models of cognitive control. *Current Opinion in Neurobiology*, 20(2):257 – 261, 2010. (page 141)

A. M. Owen, N. J. Herrod, D. K. Menon, J. C. Clark, S. P. M. J. Downey, T. A. Carpenter, P. S. Minhas, F. E. Turkheimer, E. J. Williams, E. J. Williams, T. W. Robbins, B. J. Sahakian, M. Petrides, and J. D. Pickard. Redefining the functional organization of working memory processes within human lateral prefrontal cortex. *European Journal of Neuroscience*, 11(2):567–574, 1999. (page 36)

M. G. Packard and B. J. Knowlton. Learning and memory functions of the basal ganglia. *Annual Review of Neuroscience*, 25:563–593, 2002. (page 39)

M. G. Packard and J. L. McGaughy. Inactivation of hippocampus or caudate nucleus with lidocaine differentially affects expression of place and response learning. *Neurobiology of Learning and Memory*, 65:65–72, 1996. (page 119)

M. G. Packard, R. Hirsh, and N. White. Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience*, 9(5):1465–1472, 1989. (pages 39 and 119)

G. Pagnoni, C. F. Zink, P. R. Montague, and G. S. Berns. Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, 5(2):97–98, 2002. (page 40)

A. Parker and D. Gaffan. Memory after frontal/temporal disconnection in monkeys: conditional and non-conditional tasks, unilateral and bilateral frontal lesions. *Neuropsychologia*, 36(3):259 – 271, 1998. (page 37)

R. Parr and S. Russell. Reinforcement learning with hierarchies of machines. In *Advances in Neural Information Processing Systems*, 1998. (page 78)

H. B. Parthasarathy, J. D. Schall, and A. M. Graybiel. Distributed but convergent ordering of corticostriatal projections: Analysis of the frontal eye field and the supplementary eye field in the macaque monkey. *Journal of Neuroscience*, 12(11):4468–4488, 1992. (page 38)

R. E. Passingham. *The frontal lobes and voluntary action*. Oxford University Press, 1993. (page 36)

E. Paulesu, C. D. Frith, and R. S. Frackowiak. The neural correlates of the verbal component of working memory. *Nature*, 1993. (page 36)

W. M. Pauli and R. C. O'Reilly. Attentional control of associative learning–A possible role of the central cholinergic system. *Brain research*, 1202:43–53, 2008. (page 141)

J. J. Pear and J. A. Legris. Shaping by automated tracking of an arbitrary operant response. *Journal of the Experimental Analysis of Behavior*, 47(2):241, 1987. (pages 23 and 25)

D. G. Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4):437–442, 1997. (page 86)

J. Penit-Soria, E. Audinat, and F. Crepel. Excitation of rat prefrontal cortical neurons by dopamine: An in vitro electrophysiological study. *Brain Research*, 425(2):263–274, 1987. (page 41)

G. B. Peterson. A day of great illumination: B. F. Skinner's discovery of shaping. *Journal of the Experimental Analysis of Behavior*, 82(3):317–328, 2004. (page 25)

M. Petrides. Lateral prefrontal cortex: architectonic and functional organization. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):781, 2005. (page 37)

M. Petrides and D. N. Pandya. Association pathways of the prefrontal cortex and functional observations. In D. T. Stuss and R. T. Knight, editors, *Principles of frontal lobe function*, pages 31–50. Oxford University Press, USA, 2002. (page 37)

J. R. Platt. Percentile reinforcement: Paradigms for experimental analysis of response shaping. In G. H. Bower, editor, *The psychology of learning and motivation: Advances in research and theory*, volume 7 of *Psychology of Learning and Motivation*, pages 271 – 296. Academic Press, 1973. (pages 22 and 25)

D. Plenz. When inhibition goes incognito: feedback interaction between spiny projection neurons in striatal function. *Trends in Neurosciences*, 26(8):436–443, 2003. (page 42)

D. Premack. The codes of man and beasts. *Behavioral and Brain Sciences*, 6: 125–167, 1983. (page 81)

M. L. Pucak, J. B. Levitt, J. S. Lund, and D. A. Lewis. Patterns of intrinsic and associational circuitry in monkey prefrontal cortex. *The Journal of Comparative Neurology*, 376(4):614–630, 1996. (page 36)

J. Quintana and J. M. Fuster. From perception to action: Temporal integrative functions of prefrontal and parietal neurons. *Cerebral Cortex*, 9(3):213, 1999. (page 36)

J. Randløv and P. Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, pages 463–471, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. (page 31)

A. S. Reber. Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 6(6):855–863, 1967. (page 85)

A. D. Redish, S. Jensen, A. Johnson, and Z. Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805, 2007. (pages 66, 74, 75, 79, and 146)

R. A. Rescorla and A. R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, pages 64–99, 1972. (pages 32 and 46)

L. B. Resnick, M. C. Wang, and J. Kaplan. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavior Analysis*, 6(4):679, 1973. (page 151)

J. R. Reynolds. Computational, behavioral, neuro-imaging methods investigating the hierarchical organization of pfc and goal-oriented behavior. NIPS workshop on hierarchical Organization of Behavior, 2007. (pages 83, 119, and 150)

J. R. Reynolds and R. C. O'Reilly. Developing PFC representations using reinforcement learning. *Cognition*, 113:281–292, 2009. (pages 33, 46, 59, and 148)

J. R. Reynolds, J. M. Zacks, and T. S. Braver. A computational model of event segmentation from perceptual prediction. *Cognitive Science*, 31(4): 613–643, 2007. (pages 74, 75, 79, and 147)

M. Rigotti, D. B. D. Rubin, X. J. Wang, and S. Fusi. Internal representation of task rules by recurrent dynamics: the importance of the diversity of neural responses. *Frontiers in Computational Neuroscience*, 4, 2010. (pages 140 and 148)

E. M. Robertson. The serial reaction time task: implicit motor skill learning? *Journal of Neuroscience*, 27(38):10073, 2007. (page 83)

A. J. Robinson and F. Fallside. The utility driven dynamic error propagation network. Technical Report CUED/F-INFENG/TR.1, Cambridge University engineering department, 1987. (page 47)

P. F. Rodriguez, A. R. Aron, and R. A. Poldrack. Ventralstriatal/nucleusaccumbens sensitivity to prediction errors during classification learning. *Human Brain Mapping*, 27(4):306–313, 2006. (page 40)

R. D. Rogers, T. C. Andrews, P. M. Grasby, D. J. Brooks, and T. W. Robbins. Contrasting cortical and subcortical activations produced by attentional-set shifting and reversal learning in humans. *Journal of Cognitive Neuroscience*, 12(1):142–162, 2000. (page 39)

D. L. T. Rohde and D. C. Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1): 67–109, 1999. (page 32)

H. E. Rosvold, A. F. Mirsky, I. Sarason, E. D. Bransome Jr, and L. H. Beck. A continous performance test of brain damage. *Journal of Consulting Psychology*, 20(5):343–350, 1956. (pages 57 and 83)

N. P. Rougier, N. P. Noelle, T. S. Braver, J. D. Cohen, and R. C. O'Reilly. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102(20), May 2005. (pages 11, 40, 141, and 150)

J. B. Rowe, I. Toni, O. Josephs, R. S. J. Frackowiak, and R. E. Passingham. The prefrontal cortex: response selection or maintenance within working memory? *Science*, 288(5471):1656, 2000. (page 83)

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Parallel Distributed Processing*, volume 1, chapter Learning internal representations by back-propagating errors. The MIT Press, 1986. (pages 47 and 147)

G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical report, Cambridge University, Engineering, 1994. (page 31)

K. Sakai, J. B. Rowe, and R. E. Passingham. Active maintenance in prefrontal area 46 creates distractor-resistant memory. *Nature Neuroscience*, 5 (5):479–484, 2002. (page 36)

L. M. Saksida, S. M. Raymond, and D. S. Touretzky. Shaping robot behavior using principles from instrumental conditioning. *Robotics and Autonomous Systems*, 22(3):231–249, 1997. (pages 12 and 32)

A. C. Savine and T. S. Braver. Motivated cognitive control: Reward incentives modulate preparatory neural activity during task-switching. *Journal of Neuroscience*, 30(31):10294, 2010. (page 118)

T. Sawaguchi and P. S. Goldman-Rakic. D1 dopamine receptors in prefrontal cortex: involvement in working memory. *Science*, 251(4996):947, 1991. (page 41)

A. Schäfer and H. Zimmermann. Recurrent neural networks are universal approximators. *Artificial Neural Networks–ICANN 2006*, pages 632–640, 2006. (page 47)

J. Schmidhuber. A fixed size storage o(n^3) time complexity learning algorithm for fully recurrent continually running networks. *Neural Computation*, 4(2):243 – 248, 1992. (page 47)

J. S. Schneider. Basal ganglia-motor influences: Role of sensory gating. In J. S. Schneider and T. I. Lidskey, editors, *Basal Ganglia and Behaviour: Sensory Aspects of Motor Functioning*, pages 103–121. Hans Huber, 1987. (pages 40 and 42)

W. Schultz, P. Apicella, and T. Ljungberg. Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, 13(3):900 – 913, 1993. (pages 41 and 43)

W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997. (pages 39 and 43)

B. Schwartz and S. Robbins. *Psychology of Learning & Behavior*. Norton, 1995. (page 23)

C. A. Seger. The Basal Ganglia in Human Learning. *The Neuroscientist*, 12 (4):285–290, 2006. (page 40)

O. G. Selfridge, R. S. Sutton, and A. G. Barto. Training and tracking in robotics. In *Proceedings of the 9th international joint conference on Artificial intelligence - Volume 1*, pages 670–672, San Francisco, CA, USA, 1985. Morgan Kaufmann Publishers Inc. (page 30)

D. Servan-Schreiber and J. D. Cohen. A neural network model of catecholamine modulation of behavior. *Psychiatric Annals*, 22(3):125–130, 1992. (pages 40 and 41)

D. Servan-Schreiber, J. D. Cohen, and S. Steingard. Schizophrenic deficits in the processing of context: A test of a theoretical model. *General Psychiatry*, 53(12):1105–1112, 1996. (page 83)

T. Shallice. Specific impairments of planning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 298(1089):199–209, 1982. (page 35)

T. Shallice and P. W. Burgess. Deficits in strategy application following frontal lobe damage in man. *Brain*, 114(2):727–741, 1991. (page 55)

N. Shea, K. Krug, and P. N. Tobler. Conceptual representations in goal-directed decision making. *Cognitive, Affective, & Behavioral Neuroscience*, 8 (4):418, 2008. (page 123)

F. D. Sheffield. Relation between classical conditioning and instrumental learning. In W. F. Prokasy, editor, *Classical Conditioning: A Symposium*, pages 302–322. Appleton-Century-Crofts, 1965. (page 80)

K. Shima, M. Isoda, H. Mushiake, and J. Tanji. Categorization of behavioural sequences in the prefrontal cortex. *Nature*, 445:315–318, 2007. (pages 25 and 123)

T. J. Shuell. Cognitive conceptions of learning. *Review of Educational Research*, 56(4):411, 1986. (page 10)

H. T. Siegelmann and E. D. Sontag. On the computational power of neural nets. *Journal of Computer and System Sciences*, 50(1):132 – 150, 1995. (page 13)

S. P. Singh. Transfer of learning by composing solutions of elemental sequential tasks. *Machine Learning*, 8:323–339, 1992. (pages 12, 29, 78, and 146)

B. F. Skinner. *Science and human behavior*. Colliler-Macmillian, 1953. (page 12)

B. F. Skinner. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century-Crofts, 1938. (pages 12 and 22)

S. A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22, 1996. (page 119)

S. A. Sloman and D. E. Rumelhart. Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, and S. R., editors, *From Learning Theory to Connectionist Theory: Essays in Honor of William K. Estes*, volume 1. LEA, 1992. (page 65)

J. G. Snodgrass and M. Vanderwart. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6(2):174–215, 1980. (page 91)

I. Soltesz. *Diversity in the Neuronal Machine*. Oxford University Press, New York, 2005. (page 140)

J. E. R. Staddon. *Adaptive behavior and learning*. Cambridge University Press, 1983. (page 12)

P. D. Stokes and P. D. Balsam. Effects of reinforcing preselected approximations on the topography of the rat's bar press. *Journal of the Experimental Analysis of Behavior*, 55(2):213–231, 1991. (page 24)

J. R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, 18:643–662, 1935. (page 84)

E. V. Sullivan, D. H. Mathalon, R. B. Zipursky, Z. Kersteen-Tucker, R. T. Knight, and A. Pfefferbaum. Factors of the Wisconsin Card Sorting Test as measures of frontal-lobe function in schizophrenia and in chronic alcoholism. *Psychiatry Research*, 46(2):175–199, 1993. (page 123)

C. Summerfield and E. Koechlin. A neural representation of prior information during perceptual inference. *Neuron*, 59(2):336 – 347, 2008. (page 79)

C. Summerfield, T. Egner, M. Greene, E. Koechlin, J. Mangels, and J. Hirsch. Predictive codes for forthcoming perception in the frontal cortex. *Science*, 314(5803):1311, 2006. (page 79)

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT press, 1998. (page 52)

R. S. Sutton, D. Precup, and S. Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999. (pages 12 and 28)

J. Tanji and E. Hoshi. Behavioral planning in the prefrontal cortex. *Current Opinion in Neurobiology*, 11(2):164 – 170, 2001. (page 36)

A. E. Taylor and J. A. Saintcyr. The neuropsychology of Parkinson's disease. *Brain and Cognition*, 28(3):281 – 296, 1995. (page 39)

M. E. Taylor and P. Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning. Research*, 10:1633–1685, December 2009. (pages 12, 17, and 26)

E. Teng, L. Stefanacci, L. R. Squire, and S. M. Zola. Contrasting effects on discrimination learning after hippocampal lesions and conjoint hippocampal-caudate lesions in monkeys. *Journal of Neuroscience*, 20(10):3853–3863, 2000. (page 39)

R. Thompson, D. Oden, and S. Boysen. Language-naive chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes*, 23(1):31–43, 1997. (page 81)

E. L. Thorndike and R. S. Woodworth. The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, 8:247–261, 1901. (page 22)

S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and Autonomous Systems*, 15(1-2):25–46, 1995. (pages 33 and 149)

S. Thrun and L. Pratt. *Learning to learn*. Kluwer Academic Publishing, 1998. (pages 11 and 33)

M. T. Todd, Y. Niv, and J. D. Cohen. Learning to use working memory in partially observable environments through dopaminergic reinforcement. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1689–1696. 2009. (pages 40, 46, 50, 52, 57, 82, 115, 117, and 145)

D. S. Touretzky. Boltzcons: Dynamic symbol structures in a connectionist network. *Artificial Intelligence*, 46:5–46, 1990. (pages 11 and 124)

D. S. Touretzky and G. E. Hinton. A distributed connectionist production system. *Cognitive Science*, 12:423–466, 1988. (page 11)

A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. (page 144)

J. Urzelai, D. Floreano, M. Dorigo, and M. Colombetti. Incremental robot shaping. *Connection Science*, 10(3):341–360, 1998. (pages 28 and 29)

G. Vallar and C. Papagno. In A. D. Baddeley, M. D. Kopelman, and B. A. Wilson, editors, *Handbook of Memory Disorders*, pages 249–270. Wiley, 2002. (page 36)

E. K. Vogel, A. W. McCollough, and M. G. Machizawa. Neural measures reveal individual differences in controlling access to working memory. *Nature*, 438(7067):500–503, 2005. (page 35)

U. Wagner, S. Gais, H. Haider, R. Verleger, and J. Born. Sleep inspires insight. *Nature*, 427(6972):352–355, 2004. (page 118)

J. D. Wallis, K. C. Anderson, and E. K. Miller. Single neurons in prefrontal cortex encode abstract rules. *Nature*, 411:953–956, June 2001. (pages 36 and 124)

E. K. Warrington, V. Logue, and R. T. C. Pratt. The anatomical localisation of selective impairment of auditory verbal short-term memory. *Neuropsychologia*, 9(4):377 – 387, 1971. (page 36)

M. Watanabe. Prefrontal unit activity during associative learning in the monkey. *Experimental Brain Research*, 80(2):296–309, 1990. (pages 36 and 149)

M. Watanabe, T. Kodama, and K. Hikosaka. Increase of extracellular dopamine in primate prefrontal cortex during a working memory task. *Journal of Neurophysiology*, 78(5):2795–2798, 1997. (page 41)

C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge, UK, 1989. (page 78)

P. J. Werbos. Generalisation of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1, 1988. (page 47)

D. R. Williams and H. Williams. Auto-maintenance in the pigeon: sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior*, 12(4):511–520, 1969. (page 80)

G. V. Williams and P. S. Goldman-Rakic. Modulation of memory fields by dopamine d1 receptors in prefrontal cortex. *Nature*, 376:572–575, 1995. (page 41)

R. Williams and D. Zipser. Experimental analysis of the real-time recurrent learning algorithm. *Connection Science*, 1(1):87–111, 1989. (page 55)

R. J. Williams and J. Peng. An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, 2(4):490–501, 1990. (page 50)

R. J. Williams and D. Zipser. *Back-propagation: Theory, Architectures and Applications*, chapter Gradient-based learning algorithms for recurrent networks and their computational complexity. Erlbaum, 1992. (page 47)

R. J. Williams and D. Zipser. *Backpropagation: Theory, Architectures and Applications*, chapter Gradient-based learning algorithms for recurrent networks and their computational complexity, pages 433–486. Lawrence Erlbaum Publishers, 1995. (page 47)

D. B. Willingham and W. J. Koroshetz. Evidence for dissociable motor skills in huntington's disease patients. *Psychobiology*, 21(3):173–182, 1993. (page 42)

F. A. Wilson, S. P. Scalaidhe, and P. S. Goldman-Rakic. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science*, 260(5116):1955, 1993. (page 36)

S. Wirth, M. Yanike, L. M. Frank, A. C. Smith, E. N. Brown, and W. A. Suzuki. Single neurons in the monkey hippocampus and learning of new associations. *Science*, 300(5625):1578, 2003. (page 93)

L. Wiskott, M. J. Rasch, and G. Kempermann. A functional hypothesis for adult hippocampal neurogenesis: avoidance of catastrophic interference in the dentate gyrus. *Hippocampus*, 16(3):329–343, 2006. (page 147)

M. C. Wittrock. Learning as a generative process. *Educational Psychologist*, 11(2):87–95, 1974. (page 10)

M. C. Wittrock. The cognitive movement in instruction. *Educational Researcher*, 8(2):5–11, 1979. (page 10)

J. N. Wood and J. Grafman. Human prefrontal cortex: Processing and representational perspectives. *Nature Reviews Neuroscience*, 4:139–147, 2003. (pages 35 and 37)

H. H. Yin and B. J. Knowlton. The role of the basal ganglia in habit formation. *Nature Reviews Neuroscience*, 7(6):464–476, 2006. (page 39)

A. J. Yu and P. Dayan. Uncertainty, neuromodulation, and attention. *Neuron*, 46:681–692, 2005. (pages 74 and 79)

J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds. Event perception: A mind-brain perspective. *Psychological Bulletin*, 133(2):273–293, 2007. (page 74)

E. Zarahn, G. K. Aguirre, and M. D'Esposito. Temporal isolation of the neural correlates of spatial mnemonic processing with fmri. *Cognitive Brain Research*, 7(3):255 – 268, 1999. (page 36)

C. Zhao, W. Deng, and F. Gage. Mechanisms and functional implications of adult neurogenesis. *Cell*, 132(4):645–660, 2008. (page 80)

E. A. Zilli and M. E. Hasselmo. The influence of Markov decision process structure on the possible strategic use of working memory and episodic memory. *PloS ONE*, 3:e2756, 2008. (pages 57, 82, and 145)

D. Zipser. Recurrent network model of the neural mechanism of short-term active memory. *Neural Computation*, 3(2):179–193, 1991. (pages 35 and 40)