On the Role of Temporal Context in Human Reinforcement Learning

D i s s e r t a t i o n

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

genehmigt durch die Fakultät für Naturwissenschaften
der Otto-von-Guericke-Universität Magdeburg

von: **Dipl.-Ing. Oussama Hussein Hamid**

geb. am 27. Oktober 1968 in Kuwait

Gutachter: Prof. Jochen Braun, Ph.D.

Prof. Dr. Hansjörg Scherberger

eingereicht am: 25.01.2011

verteidigt am: 10.10.2011

For my dedicative parents, my loving wife and our dearest daughter Roba

Acknowledgements

I am deeply grateful to Jochen Braun for his invaluable guidance, assistance, and training over the past 5 years. Prof. Braun's involvement in this work literally went beyond the call of duty. In fact, it is due to him that I learned: in science one needs to ask the right questions in order to be able to give good answers.

I would also like to acknowledge the helpful input of Prof. Stefano Fusi. His unsurpassable expertise in 'attractor theory' made him our first address, whenever we sought further assistance in this regard.

Many thanks to all members of the Cognitive Biology Lab at the Otto-von-Guericke University, especially, Patrick Camilleri for his support in using the computer cluster and Nicole Albrecht for managing all the sometimes sophisticated details of my conference visits.

Throughout my Ph.D., I was funded by the Federal State of Saxony-Anhalt and the Federal Ministry of Education and Technology (BMBF). A lot of thanks to all collaborators in the projects NIMITEK and UC4, in particular, Prof. Henning Scheich for the highly inspiring discussions I frequently enjoyed with him and Prof. Andreas Wendemuth and Stefan Glüge, with whom I had a journal publication and a couple of conference papers.

I would also like to faithfully thank Prof. Günther Gademann for his guidance during my work in the Clinic of Radiotherapy. Prof. Gademann shaped my thinking and provided me with the basic foundation for later decisions.

Finally, my sincerest gratitude belongs to Michiel Smid for his unfailing encouragement. Prof. Smid's cordial mentoring during my internship at Carleton University (Ottawa) in summer 2001 and his stimulating lectures at the University of Magdeburg were the first cues to guide my steps on the long path of doing research.

Abstract

Attractor network models of associative learning provide a plausible scenario for the formation of context-dependent associations. Such models make a strong qualitative prediction for temporal context: the learning of associations encompasses not only current inputs but also reverberant 'delay activity'. This implies that learned associations will include the temporal sequence of input events, whether task-relevant or not. Indeed, learning of task-irrelevant sequence information is observed in behaving non-human primates. This thesis aims at confirming and extending the above findings to human observers in order to formulate additional constraints for attractor network models.

We investigated how temporal context affects the learning of arbitrary visuomotor associations. Human observers viewed highly distinguishable, fractal objects and learned (by trial and error) to choose for each object the one motor response (out of four possible) that is rewarded. Temporal context was introduced through the sequence of objects: some objects were consistently preceded by specific other objects, while other objects lacked this task-irrelevant but predictive context.

The results of five experiments showed that predictive context consistently and significantly accelerated associative learning. A simple model of reinforcement learning, in which three successive objects informed response selection, reproduced our behavioral results.

Our results imply that not just the representation of a current event, but also the representations of past events, are reinforced during conditional associative learning. In addition, these findings are broadly consistent with the prediction of attractor network models of associative learning and their prophecy of a persistent representation of past objects.

Zusammenfassung

Auf Attraktorennetzen basierende Modelle für assoziatives Lernen liefern einen plausiblen Erklärungsansatz für die Entstehung von kontextabhängigen Assoziationen. Solche Modelle stellen eine qualitativ starke Vorhersage hinsichtlich des zeitlichen Kontextes auf: das Lernen von Assoziationen umfasst nicht nur gegenwärtige Eingabe, sondern auch die reverberierende Verzögerungsaktivität (engl. 'delay activity'). Das impliziert, dass die zeitliche Reihenfolge der Eingabe-Ereignisse mitgelernt wird. Dabei spielt es keine Rolle, ob die darin enthaltene Information aufgabenrelevant ist oder nicht. In der Tat wurde das Lernen von aufgabenirrelevanter Sequenzinformation in Verhaltensexperimenten mit Primaten beobachtet.

Die vorliegende Arbeit setzt sich zum Ziel, die Gültigkeit dieser Befunde für menschliche Probanden zu bestätigen und die erzielten Ergebnisse zu verwenden, um zusätzliche Randbedingungen für Attraktorennetze formulieren zu können.

Für diesen Zweck untersuchten wir den Einfluß des zeitlichen Kontextes auf das Lernen von arbiträren visuomotorischen Assoziationen bei menschlichen Probanden. Unsere Versuchspersonen besichtigten Sequenzen von irregulären geometrischen Objekten mit verschiedenen Formen und Farben. Ihre Aufgabe war es, durch Versuch und Irrtum zu lernen, für jedes dieser Objekte die belohnungsrelevante motorische Antwort zu wählen (visuomotorische Assoziation). Insgesamt gab es vier mögliche motorische Antworten für jedes Objekt. Der zeitliche Kontext wurde durch die Sequenz der Objekte definiert: während einige Objekte stets denselben Vorgänger in der Sequenz hatten, fehlte dieser, zwar aufgabenirrelevante, jedoch prädiktive Kontext bei anderen Objekten.

Die Ergebnisse von fünf Experimenten zeigen, dass ein prädiktiver Kontext das assoziative Lernen sowohl konsistent als auch signifikant beschleunigt.

Ein Modell des verstärkten Lernens (engl. reinforcement learning) bildete die Verhaltensdaten unserer Probanden nach. Das Modell postulierte, dass die Selektion der motorischen Antwort von drei aufeinanderfolgenden Objekten vorhergesagt wird.

Diese Ergebnisse implizieren, dass bei kontextabhängigem assoziativem Lernen nicht nur die Repräsentation eines gegenwärtigen Ereignisses belohnt wird, sondern auch die Repräsentationen von früheren Ereignissen. Diese Resultate stimmen allgemein mit den Vorhersagen von Modellen der Attraktorennetze überein.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the early days of artificial intelligence (AI), the term 'behavioral flexibility' was used to describe an agent's ability to adapt its actions to a certain environment [11, 110]. Given that the investigated environments were usually fixed [69], behavioral flexibility denoted the ability of an *inexperienced* actor to *acquire* experience through learning in a, more or less, stationary environment.

Today, as artificial systems are becoming increasingly inspired by their biological counterparts [89] and research is considering also changing environments, 'behavioral flexibility' now connotes the ability of an already *experienced* actor to *exploit* experience in order to navigate a non-stationary environment.

1.1 Context-Dependent Learning

The way how we learn and act is characterized by a wide-ranging flexibility in adapting to both (*i*) endogenous manipulations in ourselves that are due to changes in our motivational states and to (*ii*) exogenous changes in environmental factors. This very ubiquitous flexibility stems, among other things, from the facility to distill past experiences into general rules that enhance learning and shape our future behavior. Indeed,

doing well in a complex and unstable world requires a perpetual capacity to appropriately assign different forms of behavior to a given situation. Take the cultural diversity and its role in our social life as example. While Germans might shake hands, it is not unusual for Kuwaitis to kiss one another when greeting. Thus, doing business between Germany and Kuwait may well include alternating between handshaking and kissing, if etiquette forms were to contribute to a good stroke of business[1].

Advanced mammals and primates are said to be particularly quick in learning flexible rules that extend their behavioral repertoire to unfamiliar tasks and conditions. For example, well-trained monkeys can learn large numbers of arbitrary sensorimotor mappings within a few tens of trials and re-learn the new associations, when reward contingencies are *unexpectedly* changed, in only a small number of trials [4, 10, 15, 90, 94, 95, 97, 107, 138, 144]. A major determinant of this behavioral flexibility is 'context-dependent learning'. It helps the animal adjust to changing task situations without the need of extensive re-learing procedures [29, 123]. Synonymous terms are 'context conditioning' [47], 'occasion setting' [124], 'model-based reversal learning', 'goal-directed behavior', and 'outcome re-valuation/devaluation' [5, 29, 76]. They all refer to the idea of manipulating the context, in order to discern control strategies. But what is 'context' in the first place and how does it act?

The Cambridge Advanced Learner's Dictionary defines 'context' as "the situation within which something exists or happens, and that can help explain it". Applying this definition to animal learning, we may state: if a stimulus triggers more than one response during its lifetime in a certain setting, then the context should determine which of these responses is valid at any point in time [56]. For this to be possible, two conditions have to be satisfied during acquiring the context-dependency. First, the relation between the various realizations of the context in question and the different meanings of the stimulus embedded in it should be well defined [38, 85]. Second, and

[1] Do we always know what the real reason for a good deal has been?

equally important, there should be enough opportunity (mostly in terms of time and/or number of repetitions) for the context to become associated with the desired meaning of the stimulus [60]. Given this, the context would act in much the same way as an additional cue would. Hence, the widely held view: "context is just another stimulus" [46, 75].

1.1.1 The Temporal Context Hypothesis

This thesis sheds light on a special kind of context, termed 'temporal context'. Inspired by previous works [e.g. 2, 3, 91], we define the temporal context to be the amount of reward-relevant information provided by the temporal statistics of an environment in terms of the conditional probability for an event to be preceded or followed by some other events. We hypothesize that the temporal statistics of a given environment play a fundamental, and hitherto unrecognized, role in context-dependent learning. The incidental learning of temporal sequence information does not constitute an optimal decision strategy, as it gives unwarranted weight to irrelevant cues. However, incidental learning of consistent sequence information may represent a heuristic strategy suitable for natural learning scenarios, in which the relevance of environmental cues may change and previously irrelevant cues may suddenly become vitally important. For this to be shown, the current work represents behavioral data from human observers that affirm the relevance of temporal sequence information in accelerating conditional associative learning. In addition, our findings pave the way for further investigations concerning both abstract and computational frameworks like the influential 'attractor theory' and the normative theory of 'reinforcement learning', respectively.

1.1.2 Conditional Associative Learning: the Paradigm

In the laboratory, context-dependent learning was investigated using conditional associative tasks; a learning paradigm that probes the ability of primates to learn arbitrary

sensorimotor mappings [52, 109]. Typically, the experimental design takes a set of visual stimuli from the same category and maps them randomly onto a set of motor responses. Subjects learn by trial and error which response produces the reward in the case of each stimulus (e.g., if stimulus S, then response R secures the reward). As all stimuli are potentially associated with reward, the subject cannot simply learn stimulus-reward associations. Instead, subjects must link each stimulus to the specific response that ensures the reward in each case.

Learning in this way is what has been known among researchers of 'animal learning' as the *law of effect*, which was first formulated by Edward L. Thorndike [135]. It states that learning progresses *incrementally* by strengthening positively experienced associations between environmental cues and animals' responses. This requires not only stimulus recognition and response selection, but also keeping track of (at least some of) the stimulus-response pairings already tried and the outcomes obtained. Depending on the size of the stimulus set, this may generate a considerable memory load. To be more explicit, if $\mathcal{S}$ and $\mathcal{R}$ are the sets of stimuli and motor responses, with sizes $|\mathcal{S}|$ and $|\mathcal{R}|$, respectively, then the subject has to learn a *unique* correspondence among the $|\mathcal{R}|^{|\mathcal{S}|}$ many distinct and equally probable correspondences that assign the elements of $\mathcal{S}$ to those of $\mathcal{R}$. The context could help reduce the search space, as it provides additional information, which, when taken into account, might well restrict the number of potentially relevant mappings. Yet until then, subjects have to memorize the rewarded pairings and keep on trying to find out the correct motor responses for the unrewarded ones.

When the space of visual features is extended to include further dimensions (e.g. size, shape, color, orientation) or when the number of possible motor responses increases, searching for the unique correspondence becomes even more complicated. This problem has been known among researchers of 'machine learning' as the *scaling problem*. It refers to the idea that the computational time needed to reach a stable optimal

behavior grows exponentially with increasing number of environmental states and/or available actions [8, 14, 40, 69, 72, 80, 81, 129].

1.1.3 Conditional Associative Learning and the Brain

Where in the brain does conditional associative learning take place and which network of neurons is involved in this type of task managing? It is evident that conditional associative learning incorporates a wide variety of subtasks. Animals have to identify visual objects, issue motor commands, form associations, process reward values, and whenever necessary remember all of these activities at once [4, 123, 135]. Successful managing of such a complex bundle of tasks suggests that several brain areas should engage in this kind of problem solving (Fig. 1.1). In fact, studies with behaving non-human primates reveal an extensive network of brain regions underlying conditional associative learning [21, 95, 144]. The associative link between visual object recognition, subserved by inferior temporal cortex [35, 49, 83, 128, 134], and response selection, mediated by prefrontal and premotor cortex [90, 97, 138] does not, however, appear to involve a direct interaction of these brain areas [43]. Instead, conditional associative learning seems to rely on indirect pathways through the striatum [16, 58, 59, 106] and the medial temporal lobe [17, 44, 142, 146].

With more extensive stimulus sets, conditional associative tasks are suitable also for human observers. Functional imaging studies confirm that such tasks involve a similar network of prefrontal, parietal, and striatal areas in the human brain as in the brain of non-human primates [13, 19, 45, 105].

1.2 Lingering Representation of Past Events

Attractor network models of associative learning [2, 64] predict that memories should be shaped by the order in which different events are rehearsed. Commonly, these

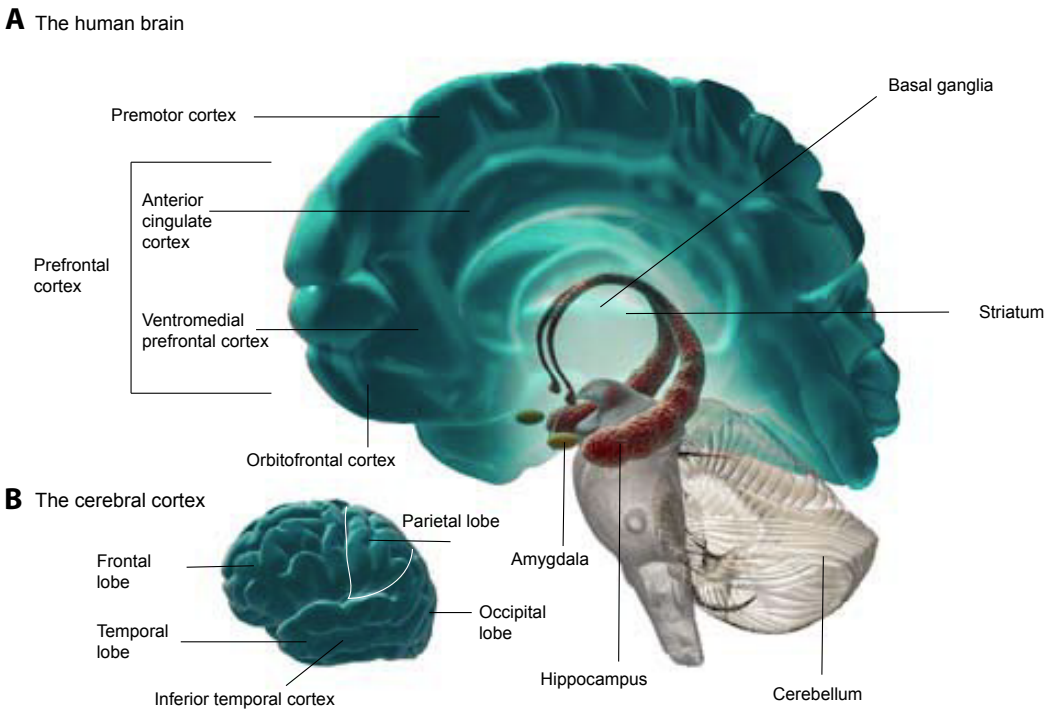

Figure 1.1: *Conditional associative learning in the human's brain.* **A**: A transparent sectional view, showing different regions of the human brain that are involved in conditional associative learning. whereas the premotor cortex contributes mainly to selecting movements in regards to the context of the action, the prefrontal cortex is responsible for executive functions. This includes action planning and decision-making. Another area is that of the basal ganlia. Beside being associated with motor control and learning, the basal ganglia are thought to also contribute to the selection of actions [29]. Located inside the medial temporal lobe, the hippocampus plays an important role in long-term memory. Finally, the inferior temporal cortex (shown in B) is crucial for visual object recognition. **B**: An exterior view of the cerebral cortex, showing the four major lobes named: frontal, parietal, temporal, and occipital. Each lobe includes many distinct functional domains. The temporal lobe, for example, has distinct regions that carry out auditory, visual, or memory functions. Adapted and modified from [65].

models assume that sufficiently strong synaptic excitations lead to the generation of self-sustained stable states of neural representations, which are manifested in the form of a persistent 'delay activity' [3]. The neural representation of an event class – its attractor state – should linger even after a triggering event has passed. Due to this reverberatory 'delay activity', events that occur consistently in a particular temporal order should eventually become subsumed under the same event class in associative memory. Importantly, it is the consistent temporal order, not mere temporal proximity, that should lead to these expanded memory representations.

1.2.1 Temporal Order Effects with Non-Human Primates

More direct evidence for an effect of temporal order on associative memory comes from electrophysiological recordings in behaving non-human primates. When monkeys are trained to perform a delayed match-to-sample task, in which they are required to determine whether a test stimulus following a delay interval matches with a sample stimulus that was presented before the delay period, neurons in the inferior temporal (IT) cortex increase their firing rates during the delay interval *selectively* for some visual stimuli [92, Fig. 1.2]. Although the monkey could perform the task also with novel visual stimuli, elevated firing rates during the delay period were observable only for highly familiar ones. This phenomenon has been known in the literature as a 'stimulus-selective delay activity' or simply 'delay activity' and was considered to be the neural correlate for forming long-term visual stimulus-stimulus associations [91]. When different sample stimuli are presented in a consistent order over successive trials, some neurons in the IT cortex develop a task-irrelevant selectivity for successive sample pairs [145].

Moreover, in monkeys trained to perform paired-associate tasks, in which they associate different objects that are presented successively, delay activity for the first object and neuronal selectivity for the pairs become evident concurrently and in the

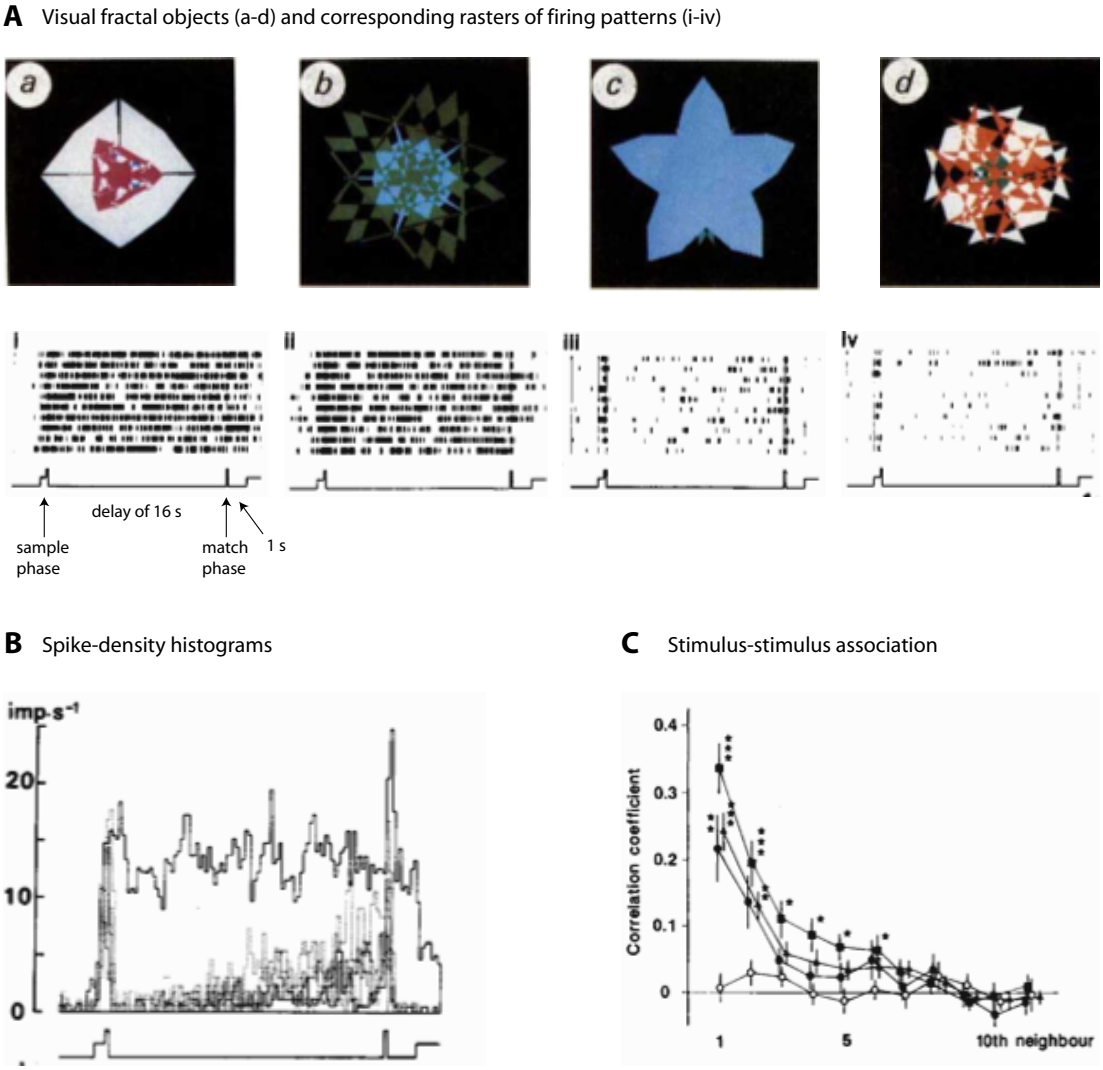

Figure 1.2: *Delay activity in a delayed match-to-sample task.* **A**(a-d): examples of the visual fractal objects used by Miyashita & Chang [92]. The raster plots **A**(i-iv) represent patterns of spike activity (dots) of *one* neuron among the 188 neurons tested in the IT cortex, with **A**(i) referring to the object presented in **A**(a) and **A**(ii-iv) corresponding to the objects shown in **A**(b-d), respectively. **B** shows spike-density histograms for the delay activity evoked by the object presented in **A**(a) and other six visual objects, for which the tested neuron was not highly selective. **C**: correlation coefficients (Kendall rank coefficients) of spike activities in a neuronal population during the delay period along the serial position number (SPN) of objects within the presented sequences. In the neurons that were tested with both learned and novel stimuli, responses to the the learned stimuli (full circles) were significantly correlated in the nearest neighbor of the SPNs, compared with the responses to the novel stimuli (empty circles). Subplots A and B adapted from [92], whereas subplot C from [91].

same neurons [121, 122] [see also 114]. These observations directly link consistent temporal order, the presence of 'delay activity', and the merging of associative memory representations.

1.2.2 Behavioral Tests with Human Observers

Behavioral results from human observers are consistent with the idea that temporal order shapes associative learning [12, 113]. For example, observers suffer in their ability to distinguish two face images after viewing image sequences in which the face identity changes as the head rotates [137]. Apparently, the correlated appearance over time leads observers to classify the two faces as the same person. Similarly, human observers come to classify two distinct objects as "similar" when they have repeatedly viewed a series of intermediate objects [112]. Importantly, the distinct objects become associated only if the intermediate objects were presented in a systematic order, starting with the most similar and ending with the most dissimilar to the initial object. Once again, it appears as if perceiving objects in a consistent temporal order would merge their representations in associative memory.

More generally, temporal order effects are well documented for serial reaction time tasks [27, 28, 116] and serial visual search tasks [24, 25], with human observers, as well as for serial button press tasks with non-human primates [62, 63].

1.3 The Attractor Framework

Exhibiting delay activity in 'working memory' tasks (like the one of the delayed match-to-sample paradigm) is not that surprising. For the animal needs (in order to perform well) to preserve the identity of the visual stimuli in 'working memory' during the delay period. Yet by using fixed sequences, the intriguing point in Miyashita's finding was the observation that the few stimuli for which a neuron was *jointly* selective, were fre-

quently *temporal* neighbors in the sequence of the visual stimuli which was presented in the training phase. Beside providing a neural correlate of associative long-term memory of pictures, this result has also confirmed the temporal order effects experimentally and corroborated the attractor network models in their predictions of persistent representations of past events in a more convincing way. Amit, Brunel & Tsodyks [2] argued that the observed delay activity is not a single neuron property, but rather an expression of the *collective* neuronal dynamics, which ultimately leads to the concretion of self-sustained stable states, termed *attractors* [see also 1].

The central tenets of attractor theory are that (*i*) the network is plastic, that is, connection strengths develop in an activity-driven, Hebbian manner and that (*ii*) associations (*e.g.*, stimulus-response pairings) are maintained as self-sustained, persistent patterns of activity which represent attractors of the neural dynamics. These tenets predict the formation of associative links whenever a set of events occurs repeatedly in a consistent temporal order [1, 3, 20, 57]. Concerning Miyashita's work, the attractor picture can be viewed as follows: every time a visual stimulus is presented, the same pattern of firing rates (delay activity) is built up across the neuronal population. Repeated presentation of a visual stimulus results in this stimulus becoming increasingly *familiar* and the corresponding pattern being more and more *stimulus-specific*. When a familiar stimulus is removed, most neurons get back to fire at their spontaneous levels, whereas some distinct ones continue firing at elevated rates in response to strong synaptic excitations by the recurrent feedback connections between the neurons. Consequently, the dynamics of the neuronal population will eventually be *attracted* into a stable state, even after the removal of the stimulus that triggered this process [1, 7, 20]. This scenario has a number of straightforward implications. Of considerable importance are the pattern-completion property and the ability of linking events in temporal order. The pattern-completion property holds that the distributed nature of neural representations of delay activity enables an attractor network to complete imperfect

patterns of delay activity, making it insensitive to possible noise. Linking events in temporal order implies in our situation that different stimulus-response pairings should not form independently, when they are rehearsed in a consistent temporal order. Instead, they should form a wider set of associative links that span successive pairings. Presumably, the corresponding attractor state would involve more neurons, be more stable, and form more rapidly. In the following we briefly explain these ideas.

1.3.1 Pattern-Completion and Noise-Insensitivty

That 'delay activity' is stimulus-specific implies that the attractor network has the ability of *pattern-completion*. Specifically, because each visual stimulus evokes a characteristic pattern of delay activity, the delay activity distribution must be referring to the neural representation of the familiar stimulus which had been seen last [145]. The distributed nature of the neural representations enables the network to store a large number of patterns in the same neural module, using the same synaptic structure [1]. If it happened that, due to noise, the delay activity of the currently presented visual stimulus is not identical (yet somehow similar) to the one, which is characteristic for that stimulus, then the network recognizes the similarity between the noisy and the noise-free (original) patterns of delay activity. As a result, the neuronal dynamics of the network will be directed to flow toward the same attractor [20]. Indeed, this ability of an attractor network to become relatively insensitive to noise has been reported for the pattern of delay activity shown by neurons in the IT cortex [3, 145].

1.3.2 Linking Events in Temporal Order

Most important for our purpose is that the attractor framework explains elegantly how associations can be formed between stimuli that are repeatedly presented in a temporal order. The basic idea is that during the delay activity, some of the neurons which are part of the current attractor will remain active, that is, they will keep firing

at higher rates until the presentation of the next visual stimulus makes the way free for the emergence of a new attractor. Meanwhile, the joint activity, which can be observed within a time window of tens of milliseconds [145], will strengthen the synaptic connections between the neuronal populations of the two attractors (the current and the previous ones). If the two stimuli were consistently presented in a fixed order, the modification of the Hebbian-manner synaptic connections, will result in both neuronal populations to have similar patterns of firing rates. This leads to the formation of associative memory [1, 2, 3, 20].

1.4 Reinforcement Learning

Understanding what the computational function of the brain is, requires to address three questions [88]. First, what is the problem the brain is trying to solve? (computational level). Second, what are the strategies it uses to solve it? (algorithmic level). Third, which networks of neurons in the brain do this and how? (implementation level). The attractor framework is a theory at the level of neural implementation. When neuronal activity is described at an appropriate level of abstraction, an attractor network model captures the collective dynamics of interacting populations of spiking neurons that is generated by recurrent connections between the neuronal populations [18]. Reinforcement learning (RL) framework provides a *complementary* approach at the level of algorithms. Specifically, whatever tasks conditional associative learning is dealing with, it is after all about making decisions. This can be defined as the process of choosing an option from a set of given alternatives. Initially, a decision maker should accumulate adequate information about candidate alternatives, then assess the corresponding values through predictions of their relative importance, then perform suitable actions, before he finally evaluates his decision in the light of the actual outcomes. RL is the *algorithmic* theory for doing this with the objective of learning optimal action

control [29, 31, 42, 69].

Though the implementation level is more concrete, as it directly deals with the structure and function of the brain, it is clear that both the computational and algorithmic levels are abstract in the sense that they can be studied and analyzed by normative treatments. Namely, in terms of developing and validating computational models and solving optimization problems. In fact, it is this point that makes RL so special and interesting for a wide spectrum of researchers, ranging from computer scientists and electrical engineers on the one side to biologists and computational neuroscientists on the other [100]. Specifically, from evolutionary perspective, animals have higher survival chances the fitter they are in terms of managing to adapt to their environments. This implies that certain behaviors turn out to be optimal solutions for problems encountered in certain situations. Studying these cases may well help theoreticians put hypotheses that can be tested computationally [68]. Moreover, observed behavior can often be better understood in the light of existing normative models. Resulting discrepancies between the predictions of a model and the real behavior can then be investigated, so as to see whether the postulated assumptions about the neural and/or informational processes do not hold or the animal is, indeed, optimizing different parameters than the ones suggested by the model.

1.4.1 Background and Inception

Historically, RL was born out of mathematical psychology and operations research [33]. Inspired by the psychological literature on Pavlovian (classical) and instrumental conditioning, Richard Sutton developed, together with Andrew Barto, algorithms for agent-based learning that later on became the core ideas for the theory of RL [132]. Parallel to their research, yet in a separate line, Dimitri Bertsekas and John Tsitsiklis, two electrical engineers working in the field of operations research, developed stochastic approximations to dynamic programming that allow a system to learn about its be-

havior through simulation (experience) and improve its performance through iterative reinforcement [11]. These lines of research marked the emergence of RL as an algorithmic theory for optimal decision making on the basis of behavior and subsequent effects [100].

Compared with other areas of machine learning that either base learning on a set of training examples (supervised learning) or seek to determine how the data are organized, by using data mining techniques (unsupervised learning), RL stands out in that it directly addresses the 'critical' question of how to optimize a policy, so as to be able to infer which of the agent's past actions led to the putative success or failure. This problem is well known in the literature as the *credit assignment problem* [132]. It comes about in situations where actions have a far reaching effect or when the outcomes depend on a sequence of actions (delayed outcomes). We encounter the *credit assignment problem* in several facets of our daily life.

In the animal world, however, the same problem can be encountered in natural scenarios like aging in case of bees or ants or even in experimental settings like searching for food in a maze [31]. Whatever the form of the *credit assignment problem* is, RL methods solve this issue by taking long-term predictions into account rather than only considering immediate outcomes when deciding which action to select in a given situation [29].

The problem of RL can be described as follows: a goal-directed agent, which might be a natural (*i.e.* biological) system or an artificial one, is interacting with an environment via sensory inputs and subjective actions. The inputs provide the agent with some information about the state of the environment. The agent responds with an action, changing the current state, before it then receives a numerical signal telling him how close it moved towards its goal or further away from it. The goal is defined in terms of some long-term measure of future utility. The simplest of conceivable measures is the cumulative expected future reward. Another possible one is the average

rate of acquisition net rewards, which is the discrepancy between positive and negative reinforcements [33]. Whatever the nature of the chosen measure is, the agent's task is to find a policy that optimizes it.

1.4.2 Markov Decision Processes

The decision process of an actor can be modeled as a *Markov decision process* (MDP) or as a *partially observable Markov decision process* (POMDP) – in case of inherent uncertainty regarding the state of the operating environment [69, 132]. An MDP process consists of two functions, $\mathbf{R}$ and $\mathbf{T}$, defined over two sets, $\mathcal{S}$ and $\mathcal{A}$. Specifically, the set $\mathcal{S}$ of states describes the different situations the agent encounters during operating (learning). The defined states differ in their nature according to the setting, in which learning is taking place. For example, while states in an operant box may best be referring to the existence or absence of different stimuli, a setting like a maze suggests to define states as possible locations within the maze. The set $\mathcal{A}$ of actions determines all feasible choices in every possible state. Examples of actions are selections of directions or presses on different levers. The reward function $\mathbf{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ refers to affectively important outcomes in form of real-valued (positive, negative, or null) reinforcements. Importantly, the outcomes can change either as a result of modifications in the motivational state of the decision-maker [5, 6, 32] or according to deliberate experimental manipulations [4, 29]. The transition function $\mathbf{T} : \mathcal{S} \times \mathcal{A} \rightarrow \Pi(\mathcal{S})$ sets probabilities for state transitions in such a way that a member of $\Pi(\mathcal{S})$ is a probability distribution over the set $\mathcal{S}$. Finally, the policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ is defined as a mapping from the set of states into the set of actions.

In the process described above, the environment evolves stochastically under simple discrete temporal dynamics in the following way: at time t, the environment is in state s_t (which might not be completely accessible to the agent). The agent chooses some action $a_t \in \mathcal{A}$, through which it *expects* to receive a certain reward $\hat{r}$ (expected

outcome). Whether matching with its expectation or not, the agent experiences the actual consequence of its choice (either immediately or later on) in form of a numerical reinforcement $r \in \mathbb{R}$ (actual outcome). Subsequently, the state of the environment changes into a new state s_{t+1} at the next time step. The agent needs to update its knowledge, so as to reflect its very experience with reward contingencies within the operating environment.

To denote the probability $P(s_{t+1} = s' | s_t = s, a_t = a)$ of moving from state s to state s' when taking action a, we write $\mathbf{T}(s, a, s')$. Analogously, we refer to the probability $P(r_t = r | s_t = s, a_t = a)$ for receiving a reward at state s_t when taking action a by $\mathbf{R}(s, a, r)$. The fact that the probability of a state transition depends only on the current state, rather than the whole history of the environment is known as the *Markov property*. Its importance in RL derives from the simplicity it provides in formalizing the reward and transition functions as functions of the current state rather than the entire history. This implies a clear computational advantage, because we need to remember and work with *only* the parameters related to the current state, which is definitely easier than dealing with all previous states of the environment [86].

1.4.3 Model-Free and Model-Based RL

Based on their optimization philosophy, RL methods can be sorted into two main classes: *model-free* and *model-based* methods [33]. Synonymous terms are *direct* and *indirect* adaptive control, respectively [69].

Though both use experience, model-free RL methods assume no prior knowledge of the environment but learn a state-action value function, termed as 'value function' [30]. Starting from a given state, the agent distills advantageous actions into an optimal policy by 'caching' actually observed information about the long-term rewarding potencies of the probed actions. In terms of computational effort, this approach represents a simple way to exploit experience, as the model needs only to learn one or two simple

quantities (state/action values). However, it is statistically less efficient, because the cached information is stored as a scalar quantity without connecting outcomes to their direct causes in a distinguishable manner. Consequently, the model's performance is most likely to suffer from two shortcomings. First, the model cannot (later on) extricate insights about rewards or transitions from the cached value. Second, the cached information intermixes previous estimates or beliefs about state values regadless their sometimes erroneous valence. As a result, model-free RL methods lack an appropriately quick adaptation to sudden changes in reward contingencies. Because of this characteristic, model-free RL was proposed as the underlying model for habitual controllers, in which actions are presumably based on habits [29]. This key characteristic links model-free RL to corticostriatal circuits involving, in particular, ventral striatum and regions of the amygdala in the human's brain [6, 32, 104].

By contrast, model-based RL captures the dynamics of the system in terms of state transition probabilities. Such probabilities can be presented as a tree connecting short-term predictions about immediate outcomes of each action in an arbitrary sequence of actions. Deciding which action is more beneficial can then be done by exploring branching sets of possible future situations. There are several 'tree search' techniques that can do this [29]. It turns out that exploiting experience as in the case of model-based RL is more efficient than in case of model-free RL for two reasons. First, it provides a more statistical reliability, especially, when storing the sometimes unrelated morsels of information. Second, and importantly, it ensures more flexibility in terms of adaptive planning, which becomes necessary when changes occur in the learning environment. Hence, model-based RL accounts best for goal-directed behavior that contain more cognitive planning. This key characteristic links model-based RL to the prefrontal cortex in the primate's brain [103].

1.4.4 The Rescorla-Wagner Model

One of the first and still most influential model-free approaches in animal learning is the Rescorla-Wagner (RW) rule. It was suggested as a model for Pavlovian conditioning; a common learning paradigm, in which animals learn to predict a reward following the presentation of a conditioned stimulus. Confirming the 'linear operator' model of Bush & Mosteller [22], which emphasized the role of 'surprise' in associative learning [39, 71], the RW model proposed that learning is driven by a prediction error, which signals the discrepancy between expected and actual outcomes. Specifically, to learn the association between a conditioned stimulus and any certain event, termed as unconditioned stimulus, one has to update expectations about the outcome in proportion to prediction error, so that across trials, the expected outcome converges to the actual outcome [118]. Formally, if the associative strength of a specific conditioned stimulus S_t in trial t was denoted by $V(S_t)$, then its value changes in the next trial $(t+1)$ according to

$$V(S_{t+1}) = V(S_t) + \alpha(S_t)\delta_t \tag{1.1}$$

where $\alpha(S_t)$ is a learning rate that can depend on the salience properties of both the conditioned and unconditioned stimuli subject to association. δ_t is a prediction error that can be computed from

$$\delta_t = r - \hat{r} \tag{1.2}$$

with

$$\hat{r} = \sum_{\tilde{S}} V(\tilde{S}_t) \tag{1.3}$$

Here, r is the actual outcome which practically limits the maximal associative strength that can be supported by the unconditioned stimulus [also known as the *asymptote of conditioning* 108]. $\hat{r}$ is the predicted outcome, which is calculated additively, by considering all conditioned stimuli that were presented in the trial. This assumption, however,

is neither the only nor always the most sensible option for combining predictions [34].

The RW model could successfully explain several characteristics of animal learning, e.g. blocking [71], overshadowing [119], and inhibitory conditioning [117]. It also proved able to predict new phenomena such as over-expectation [78]. Nevertheless, it had two shortcomings. First, it failed to account for 'secondary conditioning' [31]; a phenomenon in which a predictor of a predictor serves as a predictor. Second, and importantly, it lacked the required sensitivity to temporal aspects [100] in that it doesn't precisely handle temporal relations between conditioned and unconditioned stimuli within a trial. This is due to the fact that the RW model uses only discrete trials as temporal units, ignoring the otherwise continuous events whether in realm or even in experimental settings.

1.4.5 Temporal Difference Learning

A more time-conscious model-free method that became popular among researchers of 'machine learning' as well as those in the 'animal learning' community is the temporal difference (TD) learning algorithm. Sutton & Barto [131] introduced it as a solution to overcome the shortcomings of the Rescorla-Wagner model. As a result, the TD learning rule doesn't only take timing within a trial into account [126], but it also handle higher-order conditioning, so that problems with delayed rewards can be solved properly [31].

In this model, learning aims at optimizing the expected total future reward (starting from time t onward) by basing predictions solely on current stimuli (states) rather than considering past ones (see Markov property in section 1.4.2). Accordingly, state values can be best defined as

$$V(S_t) = E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau | S_t\right] \tag{1.4}$$

where r_τ is the reward at time τ, $E[\cdot]$ denotes the expected value, and $\gamma \in [0, 1]$ is

a discount factor that confirms the preferability of earlier rewards to delayed ones. Assuming that rewards are Bernoulli distributed with a constant probability for each state, equation (1.4) turns out to be equivalent to

$$V(S_t) = P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)V(S_{t+1}) \tag{1.5}$$

In fact, the above formula constitutes the crucial part of the TD learning rule, from which the *temporal difference prediction error* δ_t is derived. The basic idea is that the recursive relationship between consecutive state values given in equation (1.5) holds as long as the values are correctly predicted. In case of incorrect predictions, however, there will be a difference

$$\delta_t = P(r|S_t) + \gamma \sum_{S_{t+1}} P(S_{t+1}|S_t)V(S_{t+1}) - V(S_t) \tag{1.6}$$

between the left and right hand sides of equation (1.5), which can be used as a natural signal to drive learning. Yet the problem remains that the above definition of the prediction error δ_t requires knowledge of two probability distributions: the one of reward in each state $P(r|S_t)$ and that of state transitions $P(S_{t+1}|S_t)$. This stochastic information, however, can be provided by the environment incrementally. The learning system should then use experience in a way that allows to sample the missing probabilities [11, 100, 132]. However, in order to learn the true predictive state values in a model-free way, we may use the stochastic prediction error as an approximation of the true temporal difference prediction error as follows

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t) \tag{1.7}$$

where r_t is the reward delivered at time t, when in state S_t, and S_{t+1} is the next state

of the environment. As a result, state values can be updated by

$$V_{new}(S_t) = V_{old}(S_t) + \alpha(S_t)[r_t + \gamma V(S_{t+1}) - V(S_t)] \tag{1.8}$$

Applying the additivity assumption of the Rescorla-Wagner model described in equation (1.3) to the above formula, we may present the temporal difference learning rule as has been done by Sutton & Barto [131] in the following manner

$$V_{new}(S_i, t) = V_{old}(S_i, t) + \alpha(S_t) \left[r_t + \gamma \sum_{S_{k,t+1}} V_{old}(S_k, t+1) - \sum_{S_{j,t}} V_{old}(S_j, t) \right] \tag{1.9}$$

From equation (1.9) it becomes clear that the associative strength of the stimulus at time t does not restrict its predictions to the immediately forthcoming reward r_t. But rather, it transcends time limitations by considering also future predictions that are due to those stimuli, which will still be available in the next time step. Hence, using TD learning, animals can acquire the *true* predictive values of different events, even when the environment is stochastic and prior knowledge about its dynamics is not available.

1.4.6 Temporal Difference and Temporal Order

What is so interesting about the TD learning model from a neuroscientific point of view? When applied to neurobiological and behavioral data concerning the role of dopamine in reward learning and working memory, it turned out that the neurotransmitter dopamine, indeed, simulates the TD learning rule in coding the reward prediction error (Fig. 1.3). Initially, Wise, Spindler, deWit, & Gerberg [143] believed that the level of dopamine in the brain is equivalent to the reward value. Thus, blocking dopamine receptors should result in the extinction of responding by the animal, which can be, de facto, observed whenever the reward delivery is cut [48]. However, when

behaving monkeys underwent a simple instrumental or Pavlovian conditioning tasks, in which the delivery of food (reward) was *consistently* preceded by a conditioned stimulus like a tone or light, dopaminergic neurons in the ventral tegmental area (VTA) of the monkey's midbrain shifted their *reward-characteristic* phasic bursts of activity, after a number of trials, from the time of receiving the reward back to the time of perceiving the stimulus as a result of learning the underlying *temporal* association between the conditioned stimulus and the reward [82, 125]. This result challenged the 'anhedonia hypothesis' by Wise *et al.* [143], for it showed that the lack of measurable dopaminergic response was connected with acquisition rather than extinction. Drawing on this finding, Montague, Dayan & Sejnowski [93] pointed out that such pattern of dopaminergic neurons' activity conforms, in fact, exactly to the characteristics of the reward prediction error as the the TD learning rule has it [126]. Such a finding servers as a role model for the advantages of normative theories [100].

1.5 Aims of the Present Work

Historically, Miyashita's classical experiments on the learning of arbitrary visuomotor associations with non-human primates [91, 92], led to the development of attractor neural network theory of associative learning by Daniel Amit and colleagues [1, 2, 3, 7, 20, 51]. Attractor network models predict the formation of associative links whenever a set of events occurs repeatedly in a consistent temporal order. This thesis makes an attempt toward confirming and extending the Miyashita's findings, in order to formulate additional constraints for attractor network models. We introduce a novel approach to studying the effect of temporal order on associative learning with human observers. Our approach is patterned on established paradigms of conditional associative learning with non-human primates [91, 92, 121, 122, 145]. Unlike the previous studies mentioned above (section 1.2.2), the present approach does not involve

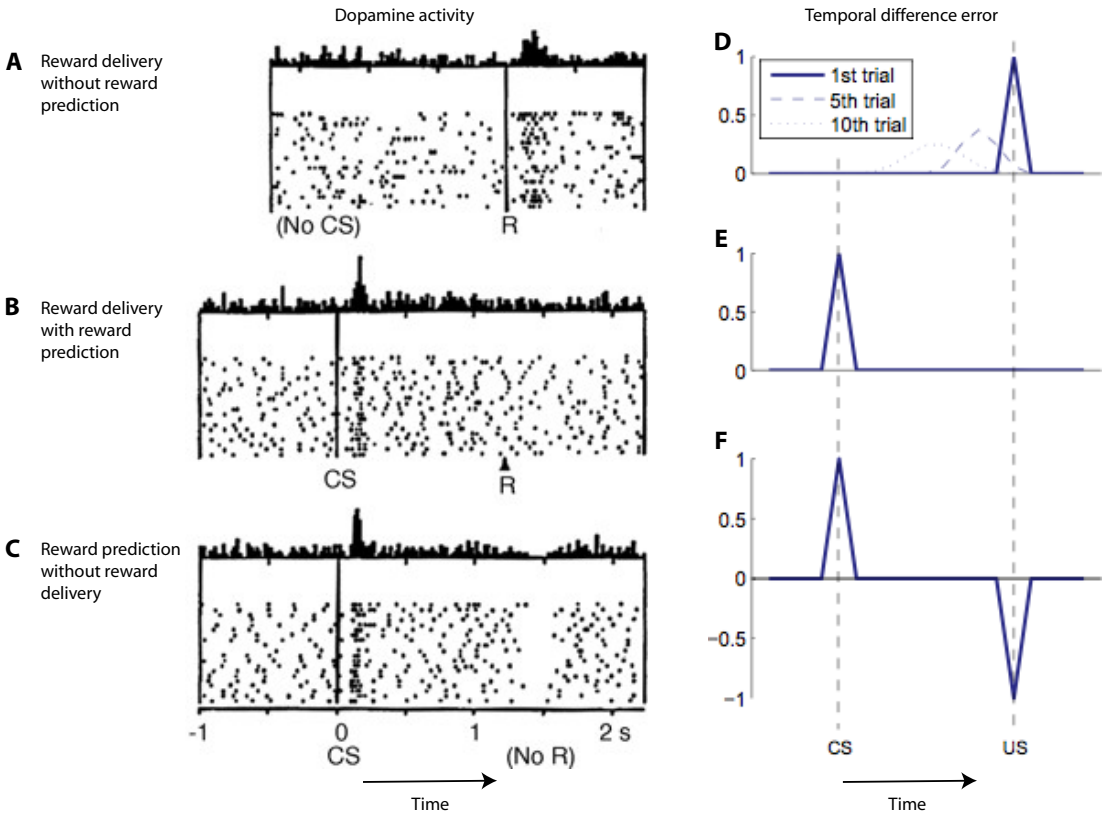

Figure 1.3: *Dopamine neurons encode the reward prediction error.* **A-C:** firing patterns of dopaminergic neurons in the ventral tegmental area of monkeys' midbrains performing instrumental conditioning task. The raster plots represent potential actions (dots) with each row referring to a trial, aligned to the time of the stimulus or the reward. The bar histograms at the top of each raster show the summed activity over the trials plotted below. Before learning **(A)**, a drop of appetitive fruit juice is delivered without the animal could have predicted it. As expected, dopamine neurons fired at elevated rates at the time of reward delivery, indicating a positive reward prediction error. However, after learning the temporal association between a predictive visual stimulus and reward **(B)**, dopamine neurons shifted their elevated firing rates from the time of reward delivery to that of the stimulus presentation. Hence, there could be no error in the prediction of reward. However, when the reward delivery was unexpectedly omitted **(C)**, firing of the dopamine neurons stopped precisely at the time where reward would have come about. **D-F:** plots of temporal difference prediction errors in a simple Pavlovian conditioning task. A tone (conditioned stimulus) is presented at random, followed 2 seconds later by a food (reward). Before learning **(D)**, the prediction error occurred at the time of reward delivery as a result of the unlearned association between the stimulus and reward. Over the course of trials **(E)**, however, the prediction error propagates back in time in correspondence to updating the values of previous time steps according to Eqn.1.8 (the plot presents trials 5 and 10 as examples). Omitting the reward unexpectedly **(F)**, generates a negative prediction error at the time reward used to be delivered at, indicating that expectation was higher than reality. Subplots (A-C) adapted from [126], whereas (D-F) from [100].

sequences of self-similar images, whether incrementally rotated [137] or morphed faces [112]. This choice was motivated by several considerations. Firstly, we wanted to stay as close as possible to the behavioral situation of the non-human primate studies in which temporal order effects were first described [91, 145]. Secondly, we wanted more freedom to manipulate temporal order than was possible with self-similar images. Thirdly, we wanted to conceal the presence of temporal order from observers, in order to minimize complications arising from cognitive strategies that often beset human studies.

Specifically, our observers viewed highly distinguishable, fractal objects and learned to select one of four possible motor responses for each object. Some objects were consistently preceded by specific other objects, while other objects lacked such a predictive temporal context (Fig. 2.2). Our aim was to keep observers engaged in the immediate task (learning visuomotor associations) and to discourage as far as possible any performance strategies relying on temporal context. For this reason, we intermixed (in most experiments) visual objects with and without temporal context and ensured that knowledge of temporal context was not necessary for accurate performance. Our results show that observers expended comparable attention and/or memory resources on objects with and without temporal context, confirming that observers applied comparable learning strategies in both cases.

In addition to experimental work, we developed several reinforcement learning models of increasing complexity [132]. Initially, a general model was devised to throw light on the way human observers solve the given task. In particular, we wanted to use the model in order to quantify the learning rate and the relevance of temporal context. This basic model, however, failed to account for our behavioral findings, for it lacked context-dependency. Consequently, we developed another model, which is (*i*) context-sensitive and (*ii*) consistent with a form of *model-free* RL. In this model, response choice is based on multiple action values, some attaching to the object of the current trial and

others attaching to objects of preceding trials. As a consequence, our model exhibits a similar dependence on temporal context as do human observers.

Although these models were not meant to capture the underlying processes (*i.e.* the plasticity and dynamics of attractor neural networks), they proved enormously helpful in developing our thinking and in shaping further experiments with non-stationary environments. In a preliminary pilot 'reversal' study, we disrupted the order of events by replacing either an individual visual object ('object reversal'), or an individual rewarded action ('action reversal'), or both of them ('combined reversal'). In either type of reversals, performance of human observers fell to chance level, contrary to the predictions of the reinforcement model. It is clear therefore, that our reinforcement model does not fully capture the way in which human observers take advantage of temporal context.

In summary, the present dissertation addressed the following:

1. We have studied the effect of temporal context on conditional associative learning.

2. Our behavioral situation is based on non-human primate paradigms but conceals the presence or absence of temporal context from human observers.

3. Our results confirm the Miyashita's findings and the predictions of attractor network models [2] in that repeated presentation of stimulus-response pairings in a consistent temporal order leads to the formation of associative links that span successive pairings.

4. We believe that this is a promising approach to testing the predictions of attractor theory of associative learning with human observers.

Chapter 2

Methods

2.1 Observers

A total of 38 female human observers (mean age: 22.5; range: 20 - 32) were recruited from the university campus. All observers reported normal or corrected-to-normal visual acuity and were naive about the purpose of the experiment. Observers completed an informed-consent form approved by the ethics committee of the university.

2.2 Apparatus and Stimuli

Highly distinguishable fractal objects with characteristic shapes and colors (Fig. 2.1) were generated in Matlab using Psychophysics Toolbox (Brainard, 1997; Pelli, 1997) with an Apple computer (Dual 2 GHn PowerPC G5; 3.5 GB SDRAM, OS x 10.4). Stimuli were displayed on a grey background of an 22 inch Iiyama color monitor with a resolution of 1900 x 1200 pixels and a frame rate of 100 Hz. The display subtended 53° at the viewing distance of 50 cm. Fractal objects were presented foveally (diameter 4°) and four response options (grey disks of diameter 4°) appeared at 4° of eccentricity above, below, to the left and to the right.

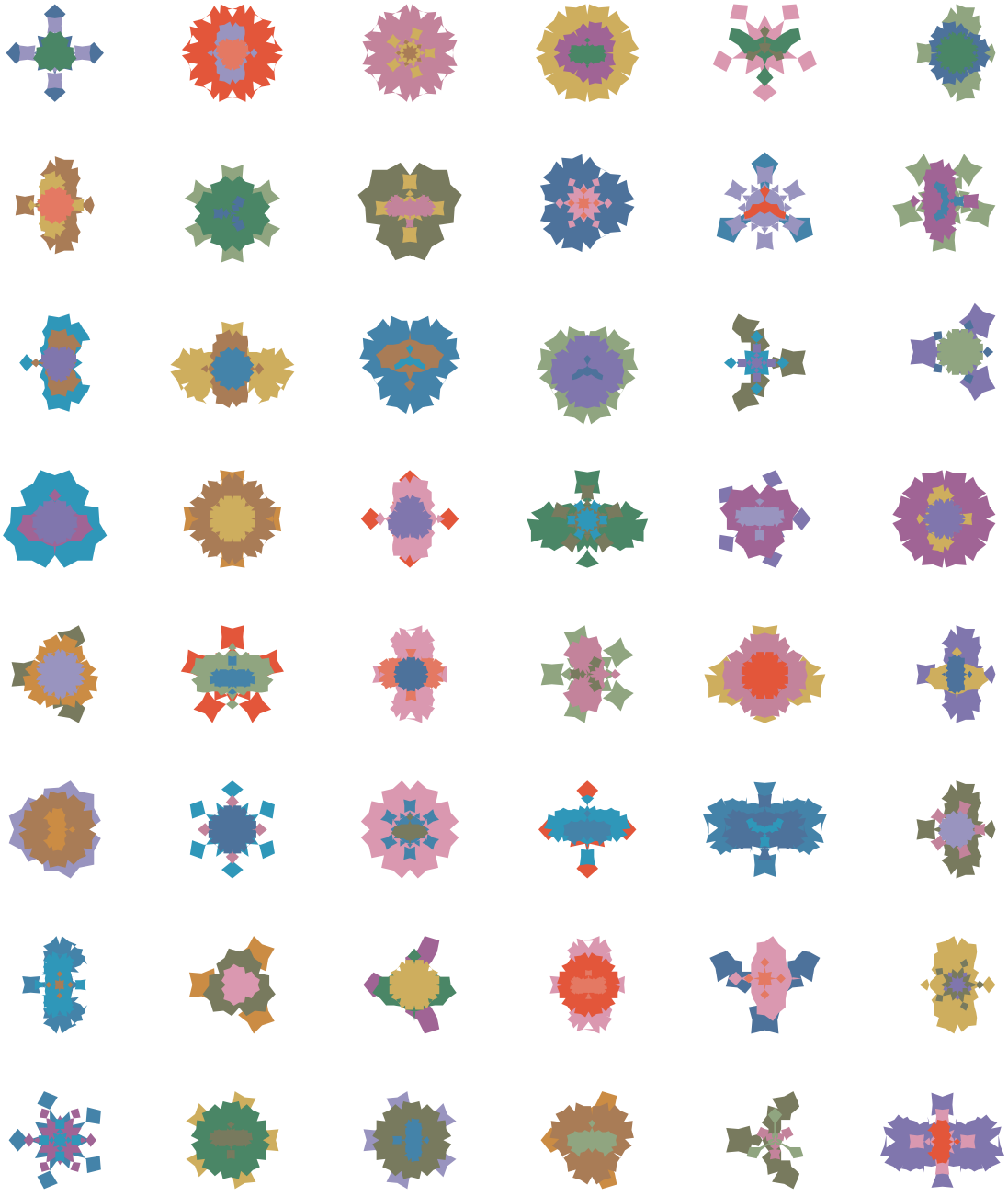

Figure 2.1: Fractal objects with characteristic shapes and colors similar to the ones used by Miyashita [91] served as visual stimuli in the present learning paradigm.

2.3 Task

Observers were instructed to learn to respond 'correctly' to each fractal object. It was explained that, for each fractal object, one of the four possible responses was 'correct', while the other three responses were 'incorrect'. Observers were told that they had to become familiar with and learn to recognize each fractal object and that they had to learn the 'correct' response of each object by trial and error. They were further told that there was no pattern or system that would enable them to predict which response a particular fractal object required. No mention of or reference to the sequence of trials and fractal objects was made.

2.4 Procedure

Each trial comprised three phases (Fig. 2.2 A): 500 ms foveal presentation of a fractal object and four response options; $500 - 2000$ ms response interval (terminated by the pressing of either $\uparrow, \rightarrow, \downarrow$, or $\leftarrow$ on the keyboard); 500 ms reinforcement (the chosen response option turned green if correct and red if incorrect). Blocks of 56 to 336 trials ('sequences') were performed without interruption. Each sequence used a new set of fractal objects, which had never before been seen by the observer.

All sequences contained 'recurring objects', each of which appeared a certain number of times (6 to 14 times) during the sequence. At least 2 trials intervened between successive recurrences of the same object. Observers typically learned the correct motor response of recurring objects (although usually the sequence was terminated before performance reached 100% correct). With sufficiently long sequences observers do reach ceiling performance.

In experiments 2 to 5, sequences also contained 'one-time objects', which appeared only once per sequence. Obviously, observers could not hope to learn the 'correct' response for such objects. However, the results suggest that observers did not distinguish

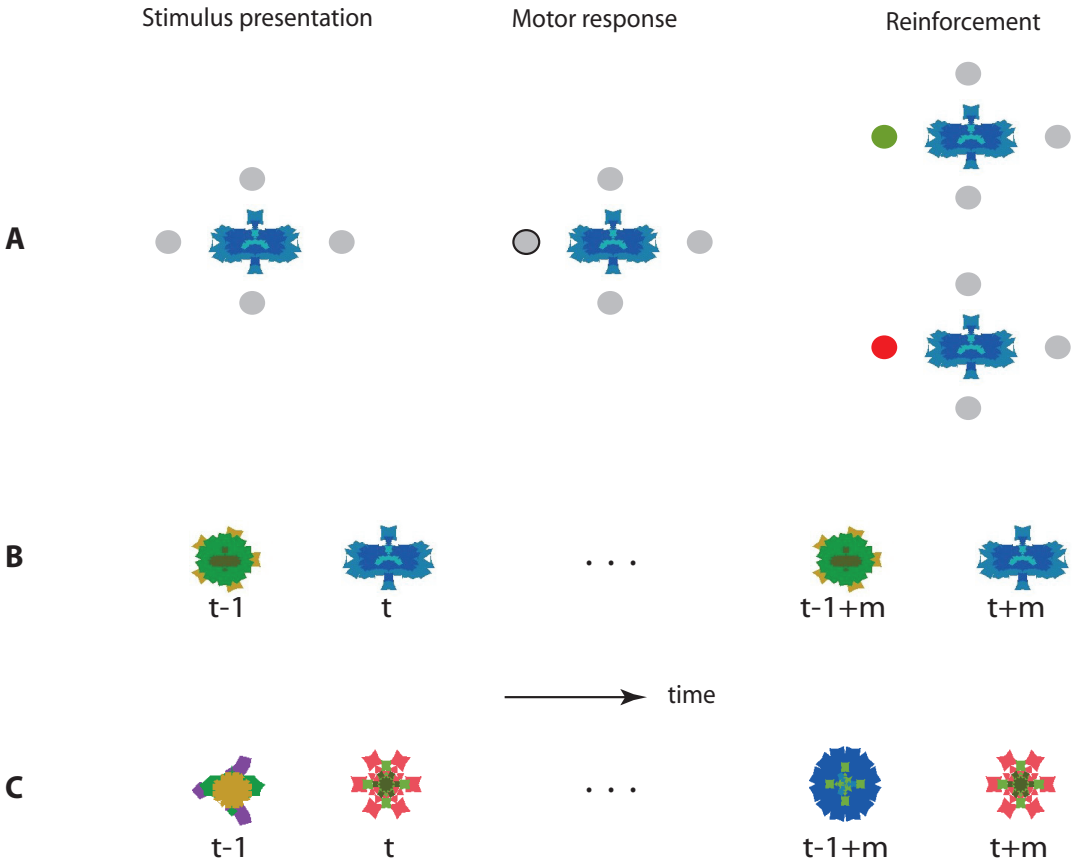

Figure 2.2: *Experimental design (schematic).* **A**: each trial comprises three phases: stimulus presentation, motor response, and reinforcement. Firstly, a fractal object appears (center), surrounded by four response options (grey discs). Secondly, the observer reacts by pressing the key that corresponds to one response option (outlined disk). Thirdly, a color change of the chosen option provides reinforcement (green if correct, red if incorrect). **B**: object sequence *with* temporal context. Target objects recur every 2 to 48 trials. Thus, successive trials always present different objects. A consistent temporal context is created by the fact that each target object (*e.g.*, trials t and $t+m$) is preceded consistently by a specific (other) object (trials $t-1$ and $t+m-1$). **C**: object sequence *without* temporal context. Each time an object appears (trials t and $t+m$), it is preceded by a different object (trials $t-1$ and $t+m-1$).

	Object type					
Experiment	A	B	C	D	E	F
1		100%	2.0%			
2	0%	100%	2.8%			
3	0%	100%	0.5%			
4			1.5%	20.3%		
5	0%	100%			0%	0%

Table 2.1: *Informativeness of temporal context.* Mutual information between predecessor object and correct response of current object, as a percentage of 2 bits (mutual information between object and correct response). See section 2.7 for details.

between recurring and one-time objects and expended comparable effort on both types of objects.

2.5 Temporal Context

Object sequences were manipulated to create a more or less predictive 'temporal context'. The current object completely determined the correct response (1 of 4 possible responses), corresponding to 2 bits of information. It is convenient to express the information provided by objects of previous trials about the correct response in the current trial as a percentage of 2 bit.

For example, the sequences in experiment 1 were either maximally deterministic or maximally random. In the deterministic sequence, each object from an earlier trial was just as informative about the correct response in the current trial as the current object (100% information). In the random sequence, objects from earlier trials carried no information about the correct response in current trials (2% information). The informativeness of the temporal contexts used in different experiments is summarized in table 2.1. The calculation of informativeness is described in the "Mutual information" section 2.7.

In experiments 2 to 5, different temporal contexts were intermixed in the same sequence. Some objects were consistently embedded in a highly informative context

(and other objects in a highly uninformative context). The types of temporal contexts used can be conveniently classified into types A to F in the following manner.

Type A: Objects were preceded by a one-time object and followed by one particular other recurring object (probability 100%). The temporal context provided by the preceding object was 0% informative in experiments 2, 3, and 5.

Type B: Objects were preceded by one particular other recurring object (probability 100%) and followed by a one-time object. The temporal context provided by the previous object was 100% informative (experiments 2, 3, and 5).

Type C: Objects were preceded (followed) by one-time objects (probability 50%) and by each of several other recurring objects (cumulative probability 50%). On average, the previous object was 2.8%, 0.5%, and 1.5% as informative as the current object (experiments 2, 3, and 4).

Type D: Objects were preceded (followed) by one-time objects (probability 50%) and by one particular other recurring object (probability 50%). On average, the previous object was 20.3% as informative as the current object (experiment 4).

Type E: Objects were preceded by a one-time object and followed by each of four other recurring objects (probability 25%). The previous object was 0% informative.

Type F: Objects were preceded by one of four other recurring objects (probability 25%) and followed a one-time object. On average, the previous object was 0% informative.

2.6 Sequences

Experiment 1: Eight fractal objects appeared seven times each, in either a deterministic or a random sequence (Fig. 2.3). Deterministic sequences were characterized

by the fact that the number of occurrences of any pair of objects, consisting of an arbitrary object and its predecessor (successor), was the same as the number of occurrences of the objects that make up this pair, that is seven times. Similarly determined pairs of objects in random sequences, however, appeared exactly once each. Both types of sequence were 56 trials long.

Experiment 2: Thirty two fractal objects were used to create sequences of 72 trials (Fig. 2.4). Eight of these objects were of the recurring kind. Four of the recurring objects formed two consistent pairs $(5, 6)$ and $(7, 8)$, each of which appeared six times in the sequence. The 'predecessor' objects (5 and 7) were termed type A and the 'successor' objects (6 and 8) type B. Four additional recurring objects were used to form twelve random pairs $(1, 2)$, $(1, 3)$, $(1, 4)$ ..., $(4, 1)$, $(4, 2)$, $(4, 3)$, each appearing once per sequence (type C). Random pairs and consistent pairs were alternated and separated by 24 one-time objects to form sequences of 72 trials.

Experiment 3: 128 fractal objects, 16 of them recurring, were used to create sequences of 336 trials (Fig. 2.5). Eight recurring objects formed four consistent pairs $(9, 10)$, $(11, 12)$, $(13, 14)$, and $(15, 16)$, each of which appeared fourteen times in the sequence. The 'predecessor' objects (odd numbers) were termed type A and the 'successor' objects (even numbers) type B. A type A object was always preceded by an one-time object. A type B object was always followed by an one-time object. Eight additional recurring objects were used to form 56 random pairs $(1, 2)$, $(1, 3)$, ..., $(1, 7)$, ..., $(8, 1)$, $(8, 2)$, ..., $(8, 7)$, each appearing once per sequence. Random pairs and consistent pairs were alternated and separated by 112 one-time objects to form sequences of 336 trials.

Experiment 4: Fifty fractal objects, ten of them recurring, were used to create sequences of 120 trials (Fig. 2.6). Five recurring objects formed twenty random

pairs $(1,2)$, $(1,3)$, (1, 4), (1, 5)..., (5, 1), (5, 2), (5, 3), (5, 4), each of which appeared twice per sequence. Objects in such pairs were termed type C objects. As before, a type C object was preceded (followed) by either another type C object or by an one-time object. Further five recurring objects were used to form five consistent pairs $(6,7)$, $(7,8)$, $(8,9)$, $(9,10)$, and $(10,6)$, each of which appeared eight times in the sequence. In contrast to earlier experiments, each object occurred in both the 'predecessor' and the 'successor' position. To mark this difference, we termed these objects type D objects. Random pairs and consistent pairs were alternated and separated by 40 one-time objects to form sequences of 120 trials.

Experiment 5: Eighty objects, sixteen of them recurring, were used to create sequences of 192 trials (Fig. 2.7). Eight recurring objects formed four consistent pairs (9, 10), $(11,12)$, $(13,14)$, and $(15,16)$, each of which appeared eight times in the sequence (type A and B). A further eight recurring objects were used to form sixteen semi-consistent pairs (1, 5), (1, 6), (1, 7), (1, 8), ..., (4, 5), (4, 6), (4, 7), (4, 8), each of which appeared twice in the sequence. The 'predecessor' objects were termed type E (1, 2, 3, 4) and the 'successor' objects were termed type F (5, 6, 7, 8). Despite being always a predecessor object, a type E object differed from a type A object in that it did not have the same successor object. Analogously, type F objects, being always successor objects, differed from type B objects in that they never had the same predecessor object. Consistent and semi-consistent pairs were alternated and separated by 64 one-time objects to form sequences of 192 trials.

2.7 Mutual Information

To convey information about the visual stimuli, reward values must be different for different stimulus-response parings. Shannon entropy [127] is a measure of variability

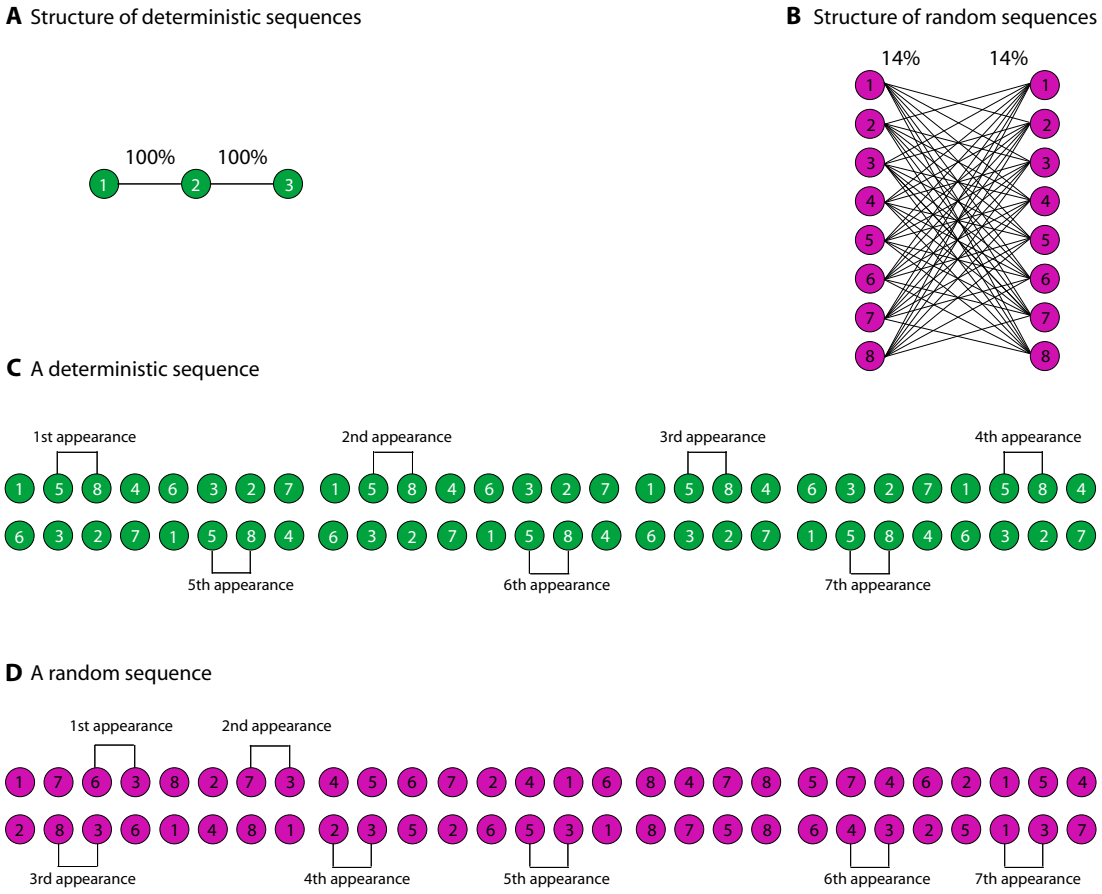

Figure 2.3: *Deterministic and random sequences (experiment 1).* Eight fractal objects appeared in either deterministic or random sequences. Both types of sequences were 56 trials long. **A**: deterministic sequences were defined by repeating a permutation of the eight objects seven times. **B**: random sequences were obtained by making each target object precede (follow) every other object exactly once (14% probability). Accordingly, target objects in both types of sequences had the same number of appearances (seven times each). Yet they differed in the number of trials between two successive appearances of the same target object (*i.e.* cyclic order). Consequently, each target object in deterministic sequences recurred every eight trials, whereas their counter parts in random sequences appeared every 3 to 14 trials. **C** and **D** illustrate examples of a deterministic and a random sequence, respectively.

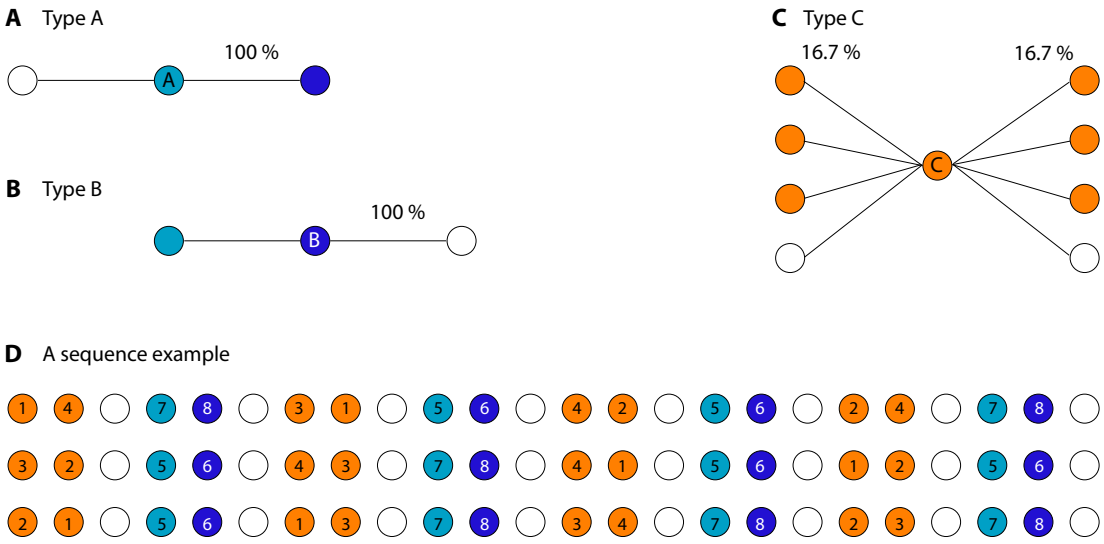

Figure 2.4: *Short mixed sequences with type A, B, and C objects (experiment 2).* Eight recurring objects (2 type A, 2 type B, and 4 type C) appeared six times each, intermixed with 24 one-time objects. **A-B:** four of the recurring objects were used to form two consistent pairs $(5, 6)$ and $(7, 8)$, each of which appeared six times in the sequence. The 'predecessor' objects (5 and 7) were termed type A and the 'successor' objects (6 and 8) type B. A type A object was always preceded by an one-time object. A type B object was always followed by an one-time object. **C:** the remaining four recurring objects 1, 2, 3, and 4 were used to form twelve random pairs $(1, 2)$, $(1, 3)$, $(1, 4)$ ..., $(4, 1)$, $(4, 2)$, $(4, 3)$, each of which appeared exactly once per trial sequence (type C). However, a type C object was preceded (followed) either by another type C object or by an one-time object. **D:** random pairs and consistent pairs were alternated and separated by 24 one-time objects to form sequences of 72 trials.

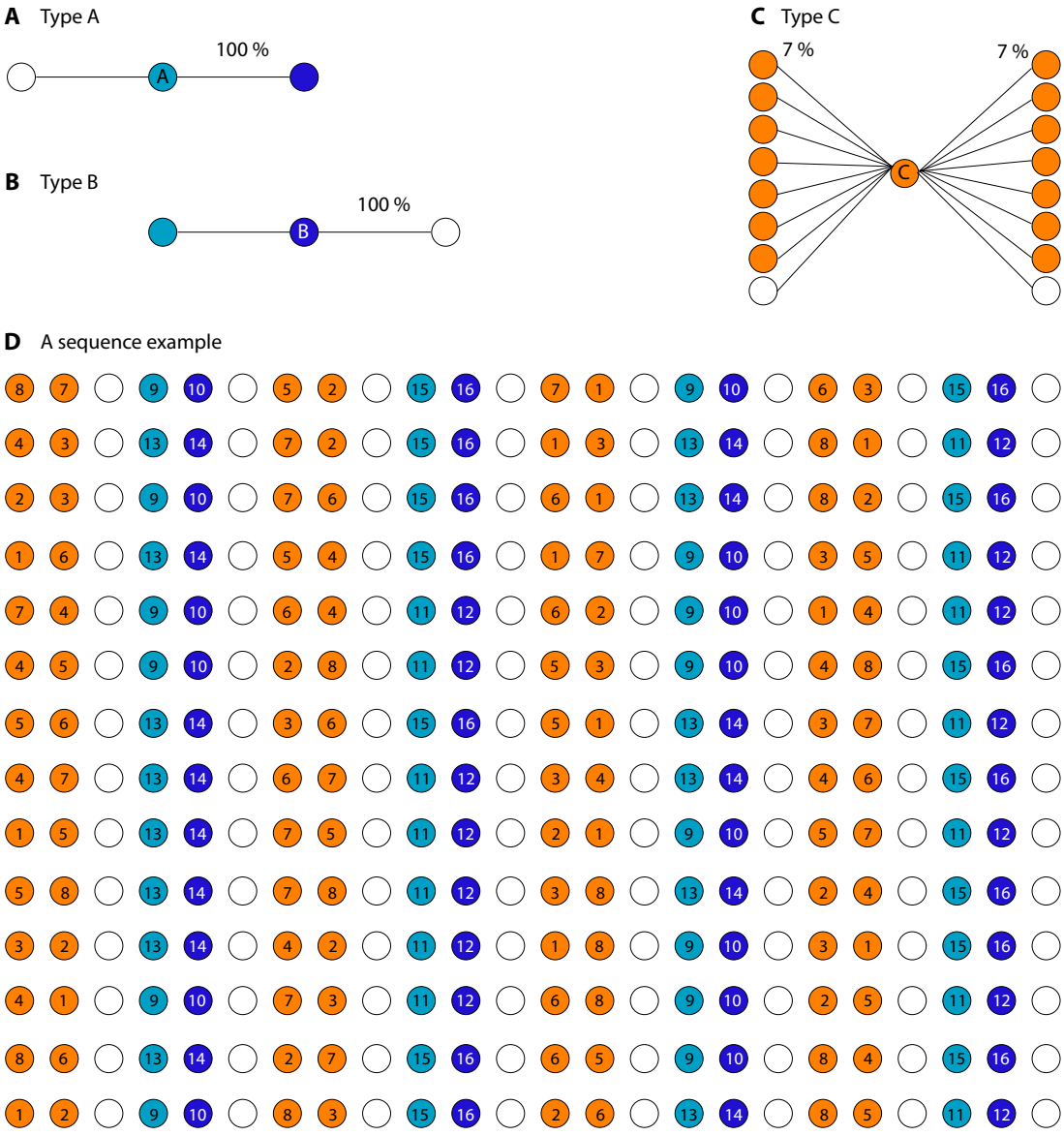

Figure 2.5: *Long mixed sequences with type A, B, and C objects (experiment 3).* **A-B**: eight recurring objects were used to form four consistent pairs $(9, 10)$, $(11, 12)$, $(13, 14)$, and $(15, 16)$, each of which appeared fourteen times in the sequence. The 'predecessor' objects (odd numbers) were termed type A and the 'successor' objects (even numbers) type B. A type A object was always preceded by an one-time object. A type B object was always followed by an one-time object. **C**: eight additional recurring objects were used to form 56 random pairs $(1, 2)$, $(1, 3)$, ..., $(1, 7)$, ..., $(8, 1)$, $(8, 2)$, ..., $(8, 7)$, each appearing once per sequence. Type C objects were preceded (followed) either by type C objects or by one-time objects. **D**: random pairs and consistent pairs were alternated and separated by 112 one-time objects to form sequences of 336 trials.

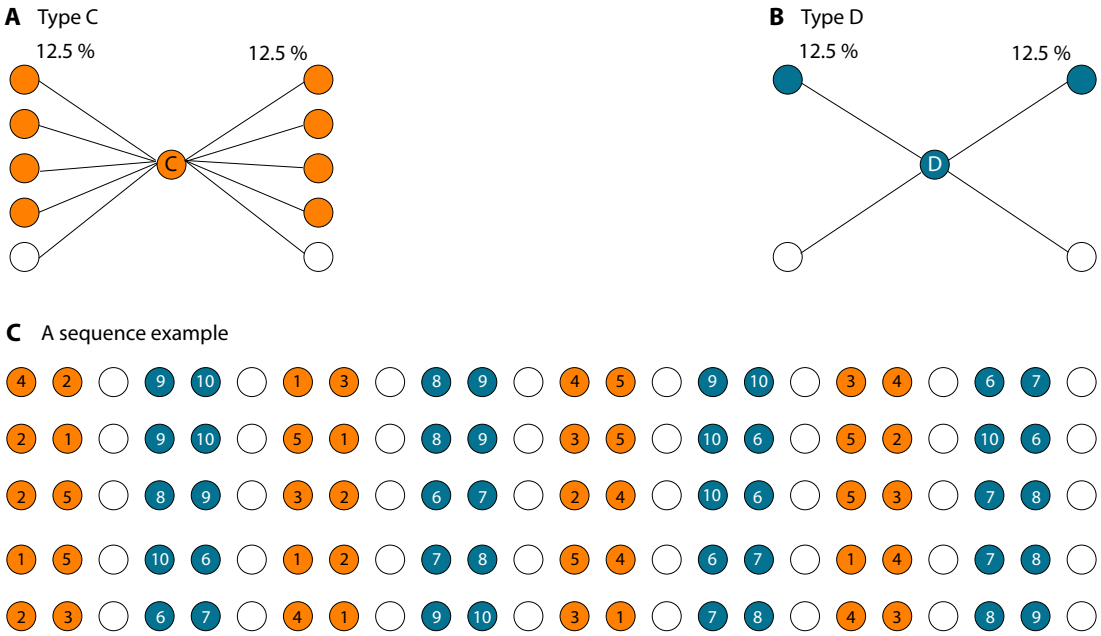

Figure 2.6: *Mixed sequences with type C and D objects (experiment 4).* **A**: five recurring objects were used to form twenty random pairs $(1, 2)$, $(1, 3)$, $(1, 4)$, $(1, 5)\ldots$, $(5, 1)$, $(5, 2)$, $(5, 3)$, $(5, 4)$, each of which appeared twice per sequence. As before, these objects were termed type C objects. A type C object was preceded (followed) by either another type C object or by an one-time object. **B**: further five recurring objects were used to form five consistent pairs $(6, 7)$, $(7, 8)$, $(8, 9)$, $(9, 10)$, and $(10, 6)$, each of which appeared eight times in the sequence. In contrast to earlier experiments, each object occurred in both the 'predecessor' and the 'successor' position. To mark this difference, we termed these objects type D objects. **C**: random pairs and consistent pairs were alternated and separated by 40 one-time objects to form sequences of 120 trials.

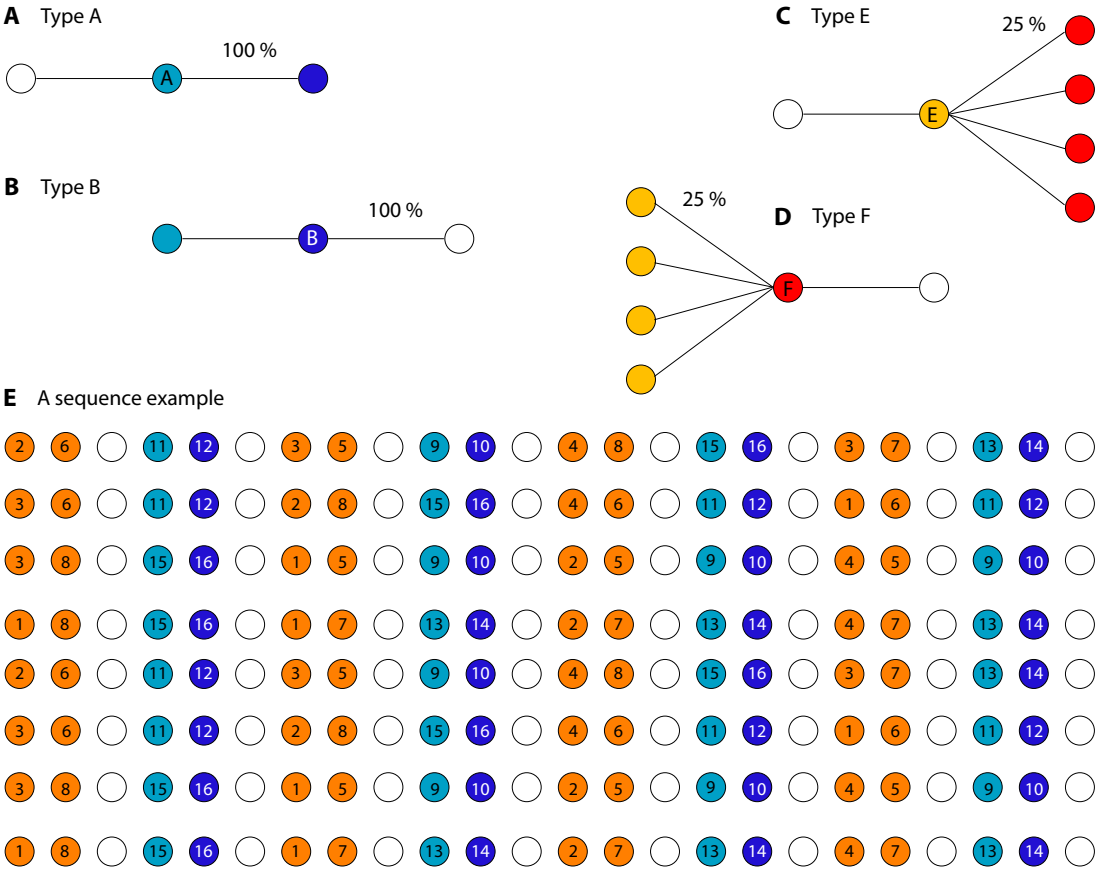

Figure 2.7: *Mixed sequences with type A, B, E, and F objects (experiment 5).* **A-B**: eight recurring objects were used to form four consistent pairs (9, 10), (11, 12), (13, 14), and (15, 16), each of which appeared eight times in the sequence (type A and B). As before, type A objects were always preceded by one-time objects, whereas type B objects were always followed by one-times objects. **C-D**: further eight recurring objects were used to form sixteen semi-consistent pairs (1, 5), (1, 6), (1, 7), (1, 8), ..., (4, 5), (4, 6), (4, 7), (4, 8), each of which appeared twice in the sequence. The 'predecessor' objects were termed type E (1, 2, 3, 4) and the 'successor' objects were termed type F (5, 6, 7, 8). **E**: consistent and semi-consistent pairs were alternated and separated by 64 one-time objects to form sequences of 192 trials.

that, by itself, does not tell us anything about the source of this variability [31]. The mutual information, however, measures how much the Shannon entropy of a random variable is reduced when we know the realization of another random variable.

We quantified the informativeness of temporal contexts in terms of mutual information. Assuming that responses are selected randomly (as is necessarily the case for unfamiliar objects), we computed the Shannon entropy H of the joint distribution of reward and motor response, conditional on the previous object

$$H = -\sum_{(m,r)} p(r_t, m_t|s_{t-1}) \log_2 p(r_t, m_t|s_{t-1}) \tag{2.1}$$

where $p(r_t, m_t|s_{t-1})$ is the joint probability of a reinforcement $r_t \in \{0, 1\}$ and a motor response $m_t \in \{1, 2, 3, 4\}$, given that a particular object s_{t-1} occurred at the preceding trial $t - 1$.

When temporal context is uninformative, a previous object does not restrict the set of possible next objects. In this case, the reward probabilities associated with the four responses are $(1/4, 1/4, 1/4, 1/4)$. The full probability matrix for the joint occurrence of a particular response and a particular motor response is then

$$\begin{pmatrix} 1/16 & 1/16 & 1/16 & 1/16 \\ 3/16 & 3/16 & 3/16 & 3/16 \end{pmatrix}$$

corresponding to an entropy of $H_{\max} = 2.8113\,\text{bit}$. When temporal context is fully informative, the presence of a previous object completely determines the next object. In this case, the reward probabilities change to $(1, 0, 0, 0)$ and the full probability matrix becomes

$$\begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

with an entropy of $H_{\min} = 2\,\text{bit}$. The mutual information between the current object

and the rewarded response is the difference between these values, or 0.8113 bit.

More generally, the informativeness of a previous object (trial $t-1$) about response-reward realization in the current trial was computed according to

$$I = \frac{H_{\max} - H}{H_{\max} - H_{\min}} \times 100\% \tag{2.2}$$

where the $H_{\max} = 2.8113$ bit and $H_{\min} = 2$ bit.

In the deterministic sequence of experiment 1, the previous object changes reward probabilities to $(1, 0, 0, 0)$ ($H = 2$ bit), whereas, in the variable sequence, the previous object changes reward probabilities to $(2/7, 2/7, 2/7, 1/7)$ ($H = 2.7953$ bit). Accordingly, in deterministic and variable sequences the previous object provides, respectively, 100% and 2.0% of the information that is provided by the current object. Conditioning on the preceding object alters the reward probabilities for type A and type B objects to $(1/4, 1/4, 1/4, 1/4)$ and $(1, 0, 0, 0)$, (entropy $H = 2.8113$ bit and $H = 2$ bit) respectively. Accordingly, the temporal context of type A and type B objects is 0% and 100%, respectively, as informative as the objects themselves. Conditioning on the predecessors of type C objects alters the average reward probabilities to $(7/24, 7/24, 7/24, 3/8)$ in experiment 2 ($H = 2.789$ bit), to $(15/56, 15/56, 15/56, 11/56)$ in experiment 3 ($H = 2.8075$ bit), and to $(9/32, 9/32, 9/32, 5/32)$ in experiment 4 ($H = 2.7992$ bit), resulting in 2.8%, 0.5%, and 1.5% informativeness. Conditioning on the predecessors of type D objects in experiment 4 alters the average reward probability to $(5/8, 1/8, 1/8, 1/8)$ with an entropy of $H = 2.6463$ bit. Thus, the predecessors are 20.3% as informative as the objects themselves. The predecessors of type E and type F objects in experiment 5 leave reward probabilities unchanged and thus are 0% informative.

Chapter 3

Behavioral Results

To ascertain whether temporal context influences the process of associative learning (or not), we conducted five behavioral experiments. In all experiments, observers learned to recognize and to classify fractal objects [91]. The objects were initially unfamiliar but highly distinguishable. For each object, observers were asked to learn the 'correct' motor response (one of four) associated with this object. After the observer's choice, the response was identified as 'correct' or 'incorrect'. Most objects recurred multiple times during the session ('recurring objects'), providing ample opportunity for learning by trial and error. Some experiments also used 'one-time objects', which appeared only once.

A trial consisted of the presentation of one object, the observer's response to that object, and reinforcement (Fig. 2.2 A). Trial sequences differed in length (56 to 336 trials) and in the number of recurring objects (8 to 16 objects), resulting in learning situations of greatly varying difficulty. Each trial sequence used new and unfamiliar objects, forcing observers to relearn the objects each time.

Pilot experiments established that human observers consistently approach ceiling performance ($P = 100\%$ correct) if the trial sequence is sufficiently long. A convenient performance measure is therefore the negative logarithm of the distance to ceiling

performance ($-\log_2(1-P)$). In terms of this measure, performance improves almost linearly with every object appearance (Fig. 3.1 A and Fig. 3.2).

The 'correct' response of each trial was determined completely by the object of that trial, which thus provided 2 bits of information. However, the object of the preceding trial was sometimes informative as well. This 'temporal context' information was redundant and, except in experiment 1, observers appeared unaware of its availability. When asked about their behavioral strategy, observers indicated consistently that they had concentrated their efforts on the current object.

The informativeness of the object in the previous trial (about the correct response in the current trial) was quantified as percentage of informativeness of the current object (see section entitled "Mutual information" in Methods). Thus, the informativeness of this temporal context ranged from 0% to 100% (Fig. 2.2 BC). Table 2.1 summarizes the informativeness of the various temporal contexts employed in experiments 1 to 5. The level of significance adopted for all the statistical comparisons reported here was set at $p < 0.05$.

3.1 Experiment 1

Eight fractal objects appeared seven times each, in either a deterministic or a random sequence (Fig. 2.3). Both types of sequence were 56 trials long. In deterministic sequences, each object was preceded (followed) seven times (100% probability) by one particular of the other seven objects. In random sequences, each object was preceded (followed) once (14% probability) by each of the seven other objects. Accordingly, the temporal context of deterministic and variable sequences was, respectively, 100% and 2% as informative about the correct response as the current object itself (see Tab. 2.1 and "Mutual information" in the Methods section).

Observers quickly understood the existence and nature of the two types of sequences

(even though the instructions had been silent on this point). Accordingly, it seemed likely that observers applied a different learning strategy in each case. The average results for 10 observers are presented in (Fig. 3.1). Post hoc t-tests revealed that learning was significantly faster in deterministic than in variable sequences ($t(239) = 2.3, p < 0.03$), exhibiting initial learning rates of 0.13 bit and 0.04 bit per appearance, respectively (average across subjects). While this difference may have been due to the disparate temporal contexts, it could also have reflected differential allocation of attentional and/or memory resources on the part of the observers.

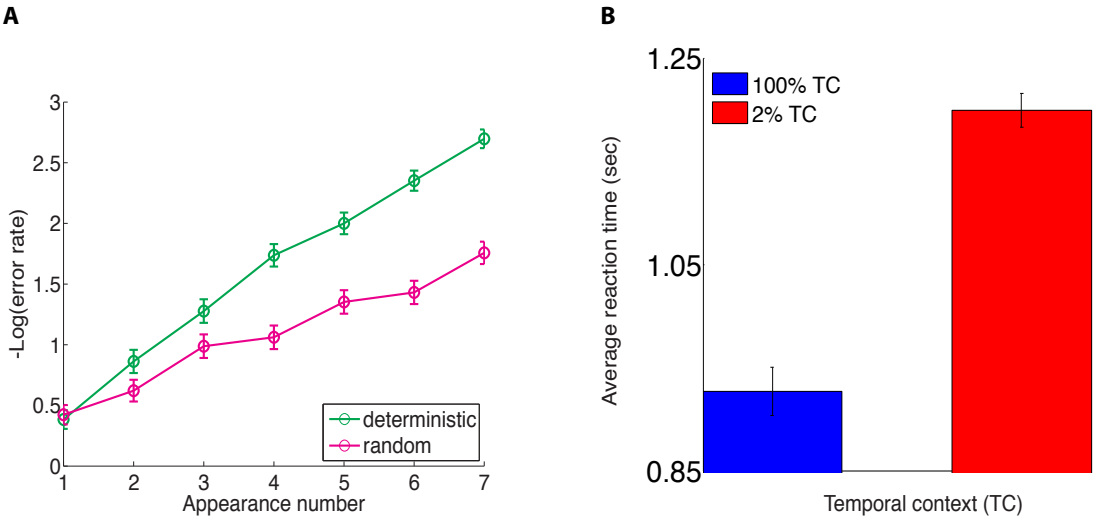

Figure 3.1: *Behavioral results* (experiment 1). **A**: average behavioral results for 10 human observers performing on two types of 56 trials long deterministic and random sequences (Fig. 2.3). Error bars refer to the 95% confidence intervals ($\alpha = 0.05$) for binomially distributed data. Observers learned the correct motor responses for objects within deterministic sequences (green curve) faster than those of objects that were presented in a random order (pink curve). **B**: average reaction times as a function of the amount of informativeness of temporal context (Tab. 2.1). Error bars show the standard deviation across experiments for each object type. Beginning with the second appearance, reaction times in experiment 1 were significantly shorter for fully predicted objects.

3.2 Experiment 2

To ascertain whether learning rate depends on the temporal context of individual objects, we created sequences that intermixed 'recurrent objects' with different temporal contexts as well as 'one-time objects'. In this situation, observers are less likely to allocate differential attentional and/or memory resources to different object types.

Eight recurring objects appeared six times each, intermixed with 24 one-time objects, in sequences of 72 trials (Fig. 2.4 D). Each of two type A recurring objects was preceded by a one-time object and followed consistently (100% probability) by one particular other recurring object (type B). Each of two type B recurring objects was consistently (100% probability) preceded by one particular other recurring object (type A) and followed by a one-time object. Each of four type C recurring objects was preceded (followed) once (16.7% probability) by each of the three other recurring objects (type C) and three times (50% probability) by a one-time object.

The temporal context of type A, B, or C objects was, respectively, 0%, 100%, and 2.8% as informative as the object itself (Tab. 2.1). The average results for 8 observers are presented in (Fig. 3.2 A). Beginning with the second appearance, learning was significantly faster for objects with more informative (type B) than with less informative (type C, type A) temporal contexts (type B vs. type A: $t(227) = 3.1, p < 0.01$; type B vs. type C: $t(227) = 2.9, p < 0.01$). The initial average rates of learning were 0.12 bit, 0.05 bit, and 0.03 bit per appearance for type B, C, and A objects, respectively.

3.3 Experiment 3

The previous experiment demonstrated that learning rate depended on the temporal context of each object in a sequence. To ascertain whether this effect would persist with a higher memory load, we conducted a similar experiment with 16 (rather than 8) recurring objects. To increase the sensitivity of the measurements, each recurring

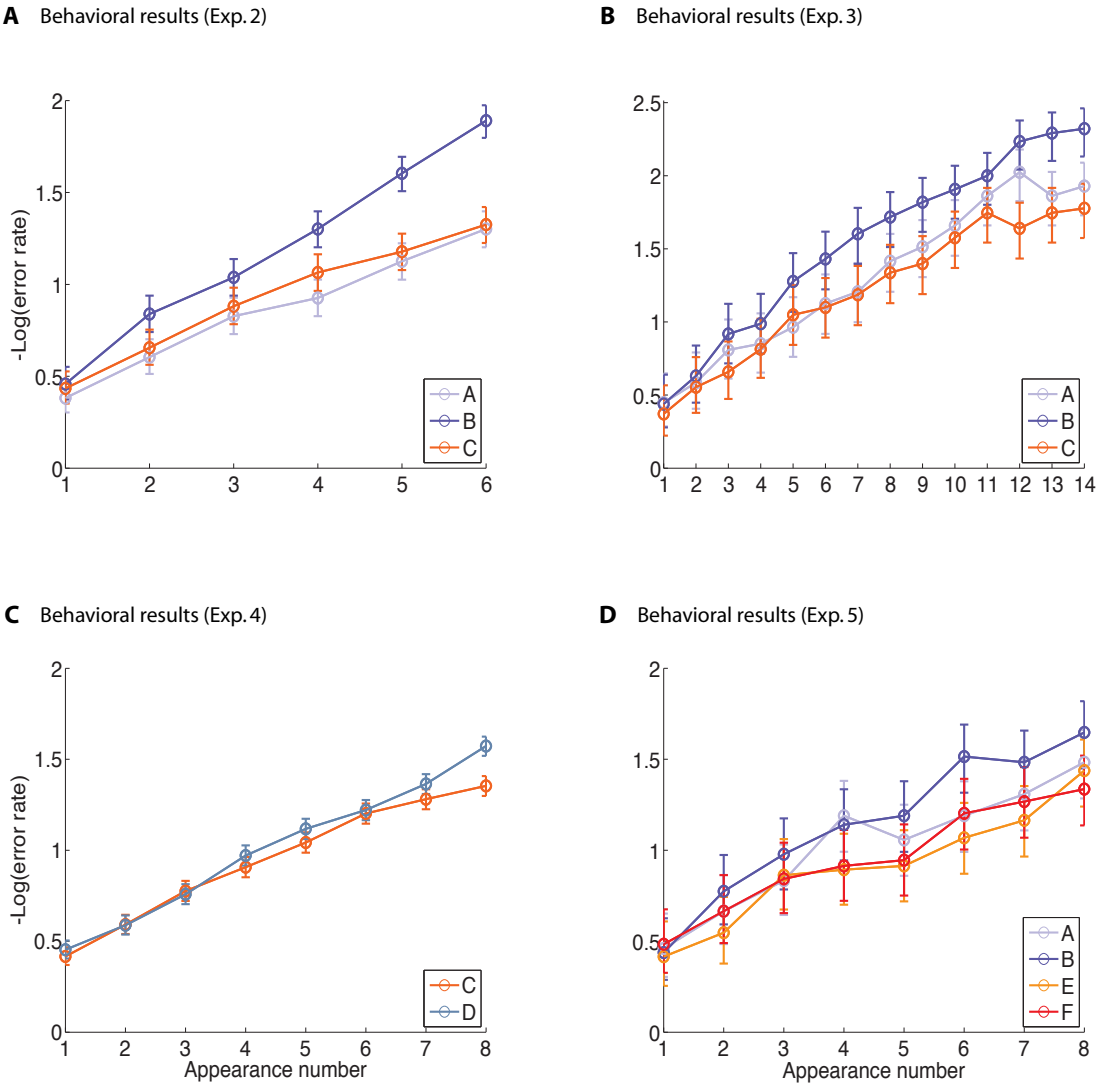

Figure 3.2: *Behavioral results* (experiments 2 to 5). Trial sequences were composed of 'recurring objects' (types A-F) distinguished by their temporal context. Error bars refer to the 95% confidence intervals ($\alpha = 0.05$) for binomially distributed data. **A**: eight recurring objects (2 type A, 2 type B, and 4 type C) appeared six times each, intermixed with 24 one-time objects. **B**: sixteen recurring objects (4 type A, 4 type B, and 8 type C) appeared 14 times each, intermixed with 112 one-time objects. **C**: ten recurring objects (5 type C and 5 type D) appeared eight times each, intermixed with 40 one-time objects. **D**: sixteen recurring objects (4 each of types A, B, E, and F) appeared 8 times each, intermixed with 64 one-time objects.

object appeared 14 (rather than 6) times.

Sixteen recurring objects appeared 14 times each, intermixed with 112 one-time objects, in sequences of 336 trials (Fig. 2.5 D). Each of four type A recurring objects was preceded by a one-time object and followed consistently (100% probability) by one particular other recurring object (type B). Each of four type B recurring objects was consistently (100% probability) preceded by one particular other recurring object (type A) and followed by a one-time object. Each of eight type C recurring objects was preceded (followed) once (7% probability) by each of the seven other recurring objects (type C) and seven times (50% probability) by a one-time object.

The temporal context of type A, B, or C objects was, respectively, 0%, 100%, and 0.5% as informative as the current object (Tab. 2.1). The results of 5 observers are summarized in (Fig. 3.2 B). Beginning with the fifth appearance, learning was significantly faster for objects with more informative (type B) than with less informative (type C, type A) temporal contexts (type B vs. type A: $t(59) = 2.2, p < 0.04$; type B vs. type C: $t(59) = 2.7, p < 0.01$). The initial average rates of learning were 0.10 bit, 0.06 bit, and 0.05 bit per appearance for type B, C, and A objects, respectively.

3.4 Experiment 4

Previous experiments compared temporal contexts that were either maximally or minimally informative. In a further experiment, we compared temporal contexts with an intermediate degree of informativeness. To this end, we presented each object in several contexts, only some of which were informative.

Ten recurring objects appeared 8 times each, intermixed with 40 one-time objects, in sequences of 120 trials (Fig. 2.6 C). Each of five type C recurring objects was preceded (followed) once (12.5% probability) by each of the four other recurring objects (type C) and four times (50% probability) by a one-time object. Each of five type D recurring

objects was preceded (followed) four times (50% probability) by one particular other recurring object (type D) and four times by a one-time object.

The temporal context of a type C or D object was, respectively, 1.5% and 20.3% as informative as the object itself (Tab. 2.1). Figure 3.2 C summarizes the results of 10 observers. Initial learning rates were comparable for type C and D objects (0.06 bit and 0.07 bit, respectively), although type D objects gained a modest advantage after further appearances. Only at the eighth (last) appearance was there a significant difference in learning between type D and type C objects ($t(689) = 2.2, p < 0.03$). The fact that observers failed to learn type D objects more rapidly than type C objects suggests that partially informative temporal contexts do not accelerate learning. Of course, it remains possible that learning would be accelerated by temporal contexts that are, say, 75% informative (*i.e.*, more than 20%, yet less than 100% informative).

3.5 Experiment 5

To allay any concern that observers might have allocated differential attention and/or memory resources to different object types, we conducted one further experiment on this point. Specifically, we presented recurrent objects in ordered pairs, some objects serving consistently as first members and others consistently as second members of these pairs. In some pairs (type A and type B objects), the first members were informative about the second members whereas, in other pairs (type E and type F objects), the first members were uninformative about the second members. If consistent object pairings had attracted additional attention/memory resources to the second member of each pairing, then this should have been true for both types of pairs, resulting in faster learning of both type B and type F objects.

Sixteen recurring objects appeared 8 times each, intermixed with 64 one-time objects, in sequences of 192 trials (Fig. 2.7 E). Each of four type A objects was preceded

by a one-time object and followed consistently (100% probability) by one particular other recurring object (type B). Each of four type B objects was preceded consistently (100% probability) by one particular other recurring object (type A) and followed by a one-time object. Each of four type E objects was preceded by a one-time object and followed twice (25% probability) by each of four other recurring objects (type F). Each of four type F objects was preceded twice (25% probability) by each of four other recurring objects (type E) and followed by a one-time object.

The temporal context of type A, B, E, or F objects was, respectively, 0%, 100%, 0%, and 0% as informative as the object itself. Figure 3.2 D summarizes the results of 5 observers. Beginning with the seventh appearance, learning was significantly faster for objects with more informative (type B) than less informative (type A, type E, and type F) temporal contexts (type B vs. type A: $t(29) = 2.24, p < 0.04$; type B vs. type E: $t(29) = 4.5, p < 0.001$; type B vs. type F: $t(29) = 2.8, p < 0.01$). The initial average rates of learning were 0.15 bit, 0.09 bit, 0.06 bit, and 0.09 bit per appearance for type B, type A, type E, and type F objects, respectively. In short, only informative temporal context led to faster learning. Merely presenting objects as consistent pairs (without the first object being informative about the second) did not accelerate learning. This failure shows conclusively that accelerated learning is due to informative temporal context, not to additional attention/memory resources.

3.6 One-Time Objects

As learning progresses, observers tend to react faster to recurring objects, whether with or without temporal context (Fig. 3.3). However, reaction times to one-time objects remained consistently slow throughout the trial sequence, suggesting that observers do try to learn (i.e., expend attentional and memory resources) even on one-time objects.

To assess the predictive value, if any, of one-time objects, we compared performance

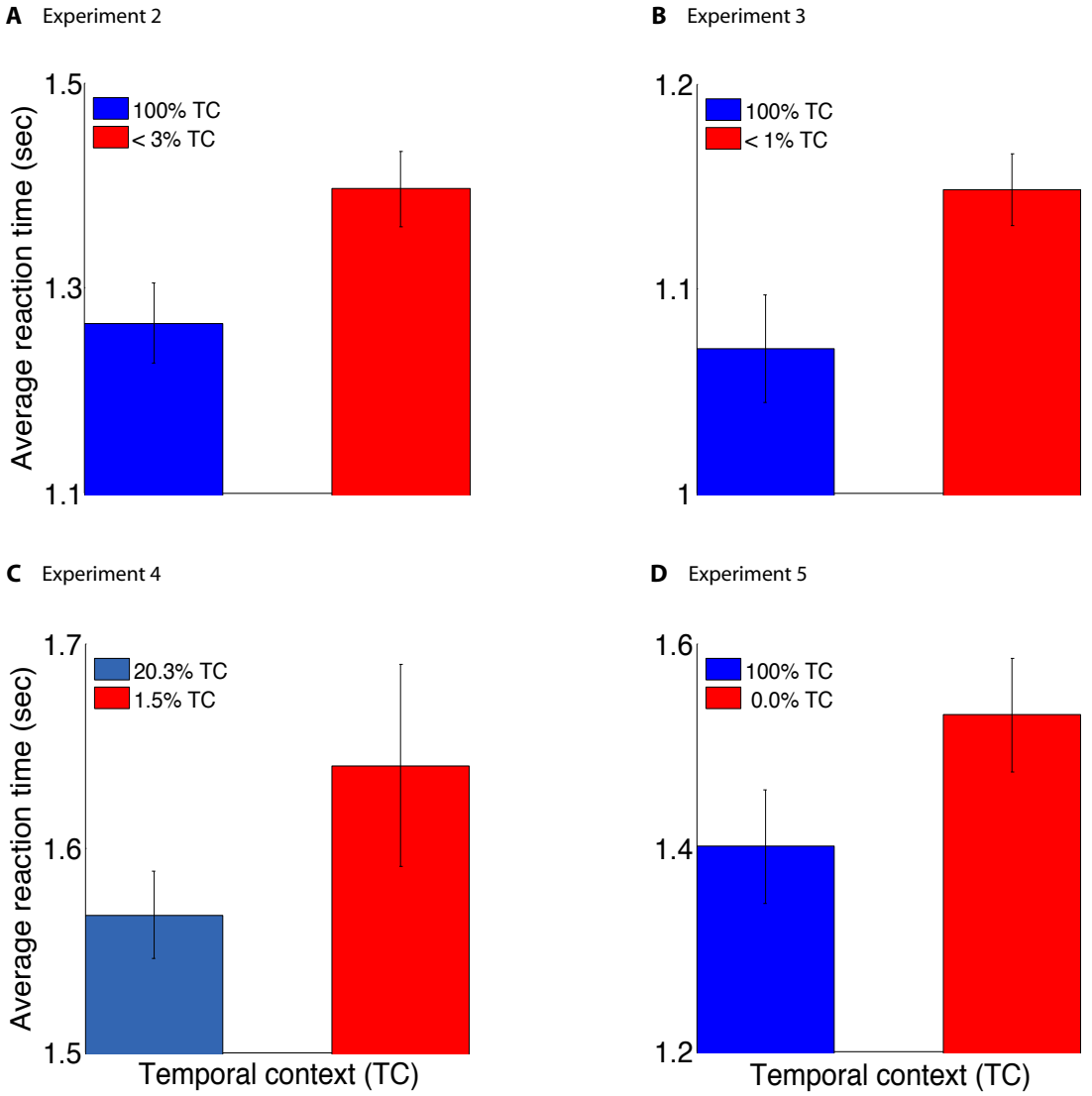

Figure 3.3: *Average reaction times due to the various degrees of temporal context* (Tab. 2.1). Error bars show the standard deviation across experiments for each object type. **A**: beginning with the third appearance, reaction times in experiment 2 were significantly shorter for fully predicted objects. **B-D**: in experiments 3 to 5, however, significant difference in reaction times due to the various amounts of temporal context became observable at the eighth, the seventh, and the third appearances, respectively.

	Object type					
Experiment	A	B	C	D	E	F
1		59%	38%			
2	35%	49%	39%			
3	23%	30%	23%			
4			12%	13%		
5	34%	37%			33%	29%

Table 3.1: *Ideal-learner-like performance.* The listed values refer to the percentage proportion of the number of blocks in which observers' responses were correct from the fourth appearance of a certain object to the last appearance of that object for each object type.

and reaction time for type C objects that followed a one-time-object and for (the identical) type C objects that followed other type C objects (experiments 2, 3, and 4). We found no significant difference in either performance or reaction time patterns between type C objects in these different contexts.

It remains possible that the (comparatively poor) performance on type A objects may have benefitted from their consistent temporal association with one-time objects. However, our sequences lacked a suitable control object so that we could not test this possibility.

3.7 Ideal-Learner-Like Performance

Since learning in our task is achieved by trial and error, any learner who never makes the same mistake twice would be an ideal learner. Drawing on this idea, we counted for each object type the number of blocks in which observers performed like an ideal learner in that they managed to respond correctly to each object from the fourth to the last appearance of that object. Though absolute correct responses within this range of trials conforms to an ideal learner's performance, yet human observers might differ from an ideal learner in the first three trials. For instance, a human observer could repeat a mistake within the first three trials or he could simply make a wrong decision after having responded correctly before. Yet in neither case can we be completely sure

that the observer, indeed, failed to recognize the object, as the probability to press the wrong button, *i. e.* make a bad decision, in the acquisition phase is higher than when observers have already learned the correct object-response associations. Therefore, we consider observers' performance from the fourth appearance of each object and term this as an 'ideal-learner-*like*' performance, if an observer continued to respond correctly till the last appearance of the considered object. Table 3.1 summarizes the result of this analysis.

3.8 Summary

An 'ideal learner' accumulates information about the correct response to a particular object at an initial average rate of 0.5 bit per appearance (see below). Human observers performed substantially less well, accumulating on average 0.09 and 0.07 bit during the initial appearance of a recurrent object in experiments 1 and 2 (memory load 8 objects), 0.07 bit in experiment 4 (10 objects), and 0.07 and 0.1 bit in experiments 3 and 5 (16 objects). These values represent learning in the absence of any temporal context provided by previous objects.

In the presence of temporal context, the accumulation of information was accelerated by 0.13 bit during the initial appearance of objects embedded in a fully predictive temporal context (Fig. 4.3 A).

Chapter 4

Computational Results

4.1 Basic Model, Insensitive to Context

A simple model for our situation is that response probabilities are modified directly such as to maximize expected reward. For each object n, four response probabilities $p_j^{(n)}$, where $j \in \{1, \cdots, 4\}$ and $\sum_j p_j^{(n)} = 1$ must be learned. When object n is observed, action k is selected, and reward $r_k \in \{0, 1\}$ is received, a suitable rule for updating response probabilities is

$$p_j^{(n)} \rightarrow \begin{cases} p_j^{(n)} + \lambda \left(\delta_{jk} - p_j^{(n)}\right) & : r_k = 1 \\ p_j^{(n)} - \mu \left(\delta_{jk} - p_k^{(n)}\right) \frac{p_j^{(n)}}{\sum_{j \neq k} p_j^{(n)}} & : r_k = 0 \end{cases} \tag{4.1}$$

where λ and μ are learning rates in the range of $[0, 1]$ and δ_{jk} is the Kronecker delta (which equals 1 if $j = k$ and 0 if $j \neq k$). This rule ensures $0 \leq p_j^{(n)} \leq 1$ and $\sum_j p_j^{(n)} = 1$. Choosing $\lambda > \mu$ makes learning faster in rewarded than in unrewarded trials. Choosing the maximal rates $\lambda = \mu = 1$ implements an 'ideal learner'.

Note that this simple model ignores temporal context and focuses on the explicit task (associating the current object with the rewarded choice). As a result, this model does not predict any dependence of learning rate on temporal context and therefore

does not account for our behavioral results.

4.2 Extended Model, Sensitive to Context

We now introduce a more elaborate model that is sensitive to temporal context. We choose an indirect actor model that responds probabilistically on the basis of reward expectations. Figure 4.1 compares the model's predictions with behavioral results by human observers.

4.2.1 Probabilistic Response

The probability of choosing response k in trial t is

$$p_k^{(t)} = \frac{\exp\left(\beta\, q_k^{(t)}\right)}{\sum_j \exp\left(\beta\, q_j^{(t)}\right)} \tag{4.2}$$

where $q_k^{(t)}$ is the reward expected from response k in trial t. The parameter β determines whether the model behaves in a more exploratory or a more exploitative manner. We use $\beta = 20$.

4.2.2 Reward Expectation

Reward expectations are based on 'action values' that have accumulated for the objects of the current trial, t, and the two previous trials, $t-1$ and $t-2$. Each object x is associated with 12 action values $m_{ij}^{(x)}$, where i indexes current, next, and after-next trials ($i \in \{0, 1, 2\}$) and j indexes the response possibilities ($j \in \{1, \cdots, 4\}$). In the case of a familiar object, action values reflect past experience as to which responses were rewarded and which unrewarded after the object in question had been observed. In the case of unfamiliar objects, all action values are initialized to 0.

Specifically, if objects n'', n', and n appeared in trials $t-2$, $t-1$, and t, respectively,

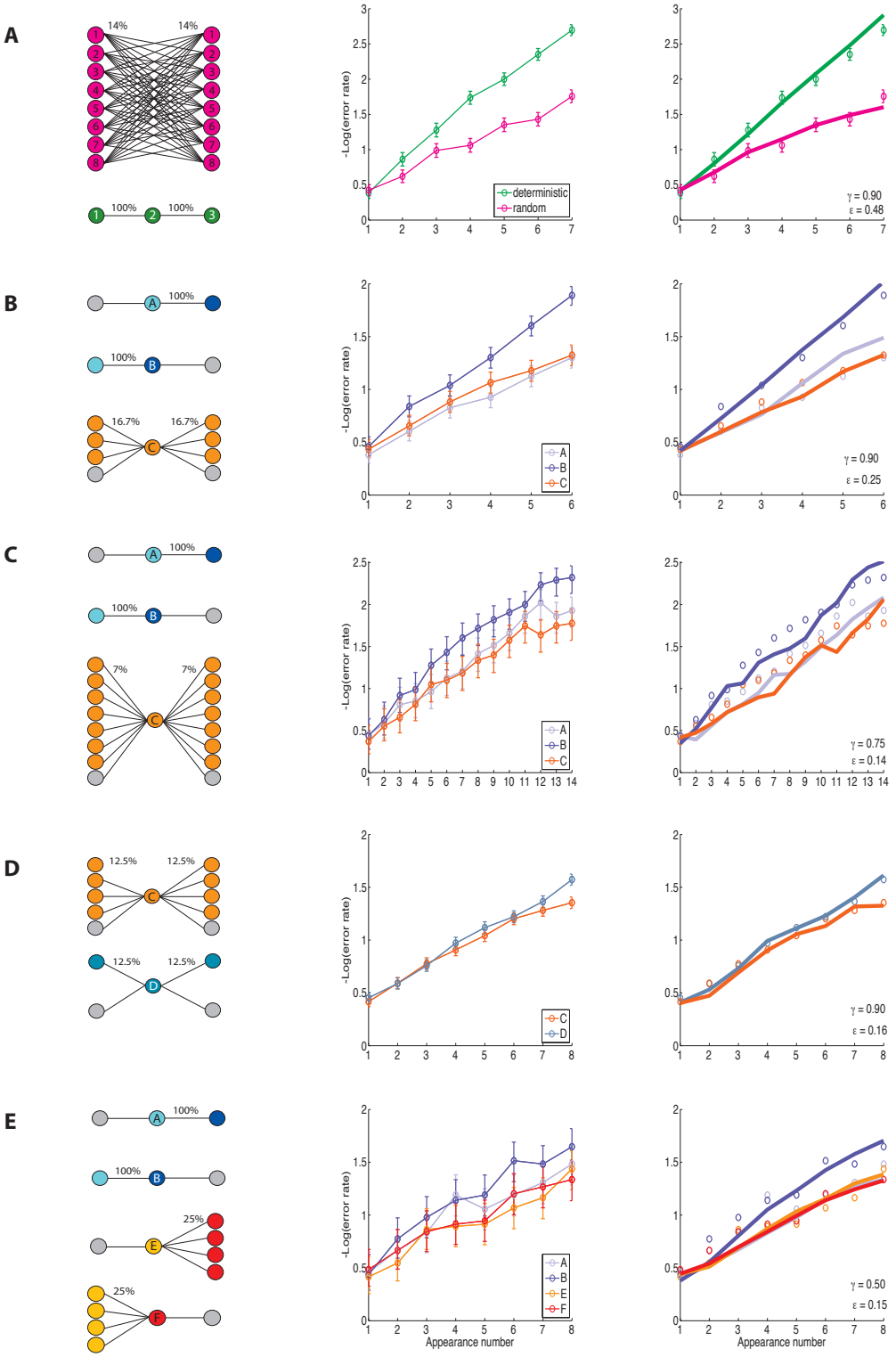

Figure 4.1: *Behavioral and modeling results.* **A-E**: for each of the experiments 1 to 5, temporal context, behavioral performance, and predicted performance are shown (left, middle, and right columns, respectively).

and if each object is recognized unambiguously, the reward expectation for response j in trial t is

$$q_j^{(t)} = m_{0j}^{(n)} + m_{1j}^{(n')} + m_{2j}^{(n'')} \tag{4.3}$$

combining action values of the current, the previous, and the before-previous objects. Temporal context determines which action values are reinforced consistently and, thus, which values come to indicate the correct response. In the absence of temporal context, only the current object's action values are reinforced consistently and thus become indicative of the correct response (Fig. 4.2). Note that the model does not assume any attenuation of past objects: current, previous, and before previous objects all contribute equally to reward expectation.

4.2.3 Action Values

Action values are reinforced by a modified Rescorla-Wagner rule [118]. If a response k receives a reward $r_k^{(t)}$ in trial t, the prediction error is

$$\delta_t = r_k^{(t)} - q_k^{(t)} \tag{4.4}$$

and the three action values $m_{0k}^{(n)}, m_{1k}^{(n')}$, and $m_{2k}^{(n'')}$ associated with action k are modified as follows:

$$m_{ik}^{(x)} \rightarrow m_{ik}^{(x)} + \epsilon\, \alpha_t^{(x)}\, \delta_t \tag{4.5}$$

where $x = n, n'$, and n'' when $i = 0, 1$, and 2 respectively, ϵ is the general learning rate, and $\alpha_t^{(x)}$ is the specific learning rate of object $x \in \{n, n', n''\}$ in trial t (see below). Action values associated with other actions $j \neq k$ remain unchanged.

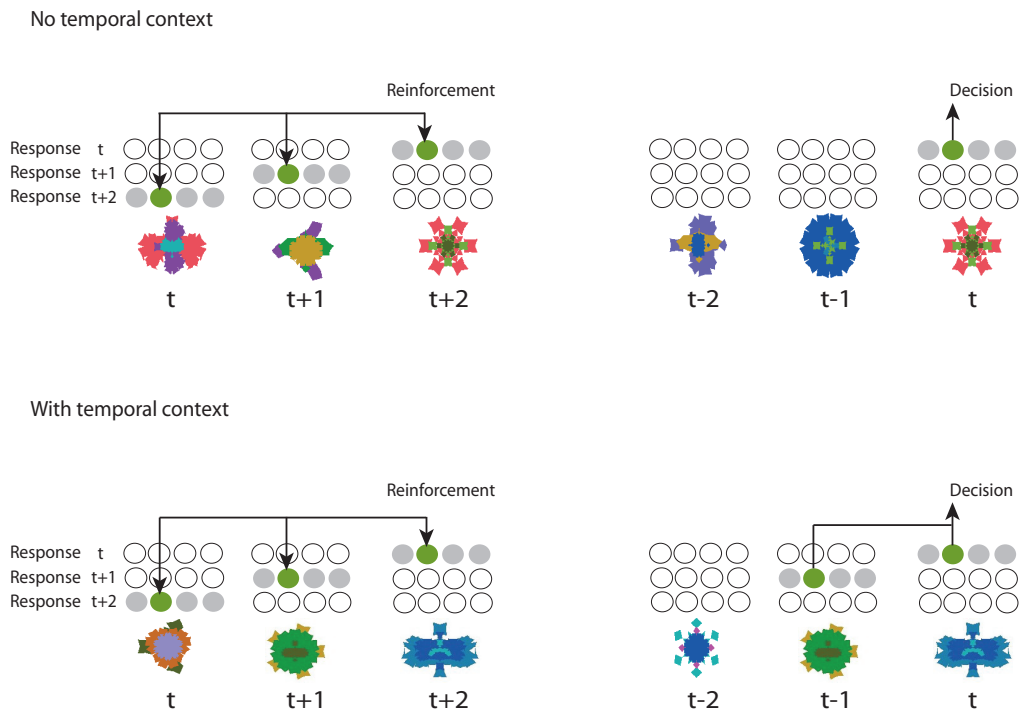

Figure 4.2: *Reinforcement of action values (schematic).* Each object is associated with 12 action values. For the object in trial t, 4 action values inform the response of the current trial t, 4 values concern the response of the next trial $t+1$, and the remaining 4 values contribute to the response of the second next trial $t+2$. Correspondingly, the response of trial t is based on 12 actions values: 4 values of the current object t, 4 values of the previous object $t-1$, and 4 values of the pre-previous object $t-2$. Temporal context determines which action values are reinforced consistently. **A:** in the absence of temporal context, only the current object's action values are reinforced consistently and come to reflect the correct choice. In this case, the decision in trial t is based on 4 action values of object t. **B:** in the presence of temporal context, both the current and the previous object's action values are reinforced consistently. Thus, the decision in trial t is based on 4 action values of object t and 4 action values of object $t-1$.

4.2.4 Recognition Parameter

Human observers sometimes fail to recognize an object they have seen before. To model this confusion about object identity, we introduce a *recognition parameter* γ, $0 \leq \gamma \leq 1$, which parametrizes the extent to which an actual object is recognized as being present. The value of γ affects learning in two ways. Firstly, it influences the reward expectation by taking into account not only the objects actually present but also all other objects. As a result, equation (4.3) becomes

$$q_j^{(t)} = M_{0j}^{(n)} + M_{1j}^{(n')} + M_{2j}^{(n'')} \tag{4.6}$$

where $M_{ij}^{(x)} = \gamma\, m_{ij}^{(x)} + \frac{1-\gamma}{N-1} \sum_{y \neq x} m_{ij}^{(y)}$ for $i \in \{0,1,2\}$ and $x \in \{n, n', n''\}$. N is the total number of objects. Secondly, $\gamma < 1$ removes some reinforcement from action values of objects actually present and distributes the reinforcement over the action values of all other objects. Accordingly, equation (4.5) modifies to

$$m_{ik}^{(y)} \rightarrow \begin{cases} m_{ik}^{(y)} + \epsilon\, \alpha_t^{(y)}\, \gamma\, \delta_t, & : y = x \\ \\ m_{ik}^{(y)} + \epsilon\, \alpha_t^{(y)}\, \frac{1-\gamma}{N-1}\, \delta_t & : y \neq x \end{cases} \tag{4.7}$$

where $i \in \{0,1,2\}$ and $x \in \{n, n', n''\}$.

The recognition parameter γ is an admittedly crude way of modeling confusion about object identity. In human observers, one might expect that recognition rates increase with every appearance of a particular object. In our model, the value of γ does not reflect this (hypothetical) improvement and remains constant throughout the sequence.

4.2.5 Specific Learning Rates

Another fundamental feature of learning that has been postulated by the Rescorla-Wagner rule is the learning rate. It reflects the extent to which current knowledge about the environment should be considered when the learner receives new information. Despite its necessity in the learning process, whether in biological or artificial systems, it has never been clear, how or why it changes [9, 41]. In general, several stimuli may be associated with rewards [34]. As to reflect how reliably a particular object is associated with the reward, we computed 'specific' learning rates using the Kalman-filter algorithm [70] as proposed by Sutton [130].

Specifically, let $\mathbf{x}^{(t)}$ be the augmented stimulus vector of trial t which comprises three components for each object $n_i \in \{n_1, \ldots, n_N\}$ (one component for each the current, the previous, and the before-previous trial). The values of $\mathbf{x}^{(t)}$ reflect the recognition parameter and differ for present and absent objects in the following manner:

$$x_j^{(t)} = \begin{cases} \gamma & : n_i \text{ present} \\ \frac{1-\gamma}{N-1} & : n_i \text{ absent} \end{cases} \tag{4.8}$$

where $j \in \{1, \ldots, 3N\}$ and $i = j \mod N$. The specific learning rate of object x_i is computed from

$$\alpha_t^{(x_i)} = \frac{\sum_i P_{ij}^{(t)} x_j^{(t)}}{1 + \sum_i \sum_j x_i^{(t)} P_{ij}^{(t)} x_j^{(t)}} \tag{4.9}$$

where $P_{ij}^{(t)}$ is a drift covariance matrix that is accumulated iteratively.

Sutton [130] evaluated several dynamic-learning-rate methods for the selection of learning rates or gain parameters during learning of stochastic time-varying linear systems. He showed that the Kalman-filter, though requires prior knowledge for an estimate of the drift covariance matrix P, indeed performs optimally well in terms of asymptotic error when compared with least-squares methods. However, in practice the drift covariance matrix is never known exactly. Thus an approximation must be used.

In order to update the drift covariance matrix $P^{(t)}$, we used the same equation as the one given in Sutton [130]:

$$P^{(t+1)} = P^{(t)} - \frac{P^{(t)}\mathbf{x}\mathbf{x}^T P^{(t)}}{1 + \mathbf{x}^T P \mathbf{x}} + I$$

where I is the identity matrix and $\mathbf{x}$ is the augmented stimulus vector. Once initialized ($P^{(0)} = I$), the drift covariance matrix $P^{(t)}$ is computed recursively and an iteration takes place as follows:

1. $A^T = \mathbf{x}^T P = [\sum_i P_{ij} x_i]^T$

2. $B = P\mathbf{x} = \sum_j P_{ij} x_j$

3. $C = \mathbf{x}^T P \mathbf{x} = \mathbf{x}^T B = \sum_i \sum_j x_i P_{ij} x_j$

4. $\alpha^{(t)} = B(1 + C)^{-1}$

5. $P^{(t+1)} = P^{(t)} - BA^T(1 + C)^{-1}$

The superscript T in A^T indicates the transpose of A and $(\cdot)^{-1}$ denotes the inverse matrix, which is the reciprocal in case of numbers.

It is important to note that we could have passed on the specific learning rates, as they barely contribute to a substantial improvement of the model's performance in the current situation. In fact, our simulations (not included here), in which identical learning rates were used, instead of differential ones (like those obtained by the Kalman filter) have shown that even in this case the model manages to capture the essence of temporal context. Yet the reason why we chose to implement this part of the model using 'specific' learning rates was to generalize this approach to non-stationary environ-

ments. Such environments are characterized by changing reward contingencies (reversal learning), which yield a change in the informativeness of objects. The Kalman filter approach specifies how the uncertainty accompanying predictions about the informativeness of the various objects change over time. Compared to the identical-learning-rates-approach it provides a better solution in that it adjusts the speed of learning according to the uncertainty of the prediction. This 'competitive' allocation of learning between the objects according to their uncertainties imply that certain predictions change more slowly, whereas comparatively uncertain predictions more quickly [34].

4.3 Model Fitting

In both basic and extended models, response choices depend on 'action values' that are learned by reinforcement. The basic model, in which action values are associated exclusively with the current object, ignores temporal context in choosing the current response. As a result, the basic model does not account for the sensitivity to temporal context exhibited by human observers. Nevertheless, the basic model provides a useful benchmark to which human performance can be compared.

With learning rates set to their maximal values of $\lambda = \mu = 1$, the basic model implements an 'ideal learner'. Its average performance increases from 25% correct on the first appearance of an object, to 50%, 75%, and 100% correct on the second, third, and fourth appearance of the object. The combined entropy of response and reward falls from 2.81 bit on the first appearance, to 2.16 bit, 1.41 bit, and 0 bit on the second, third, and fourth appearances, respectively.

In the extended model, action choices are influenced equally by three objects: the current, the previous, and the one preceding the previous object. In addition to this sensitivity to temporal context, the extended model also allows for probabilistic object recognition and employs differental learning rates that depend on the reliability of a

reward-association [130].

The extended model has two free parameters, namely, the general learning rate ϵ and the recognition parameter γ (Eqn. 4.7). The parameter β did not materially affect the results and its value was kept equal to $\beta = 20$ throughout (Eqn. 4.2).

The extended model was fit to the behavioral results in the ranges of $0 \leq \epsilon \leq 1$ and $0.25 \leq \gamma \leq 0.9$ (Fig. 4.3). The results of experiment 1 are consistent with a comparatively rapid learning rate of $\epsilon \approx 0.48$ and a near-perfect recognition probability of $\gamma \approx 0.9$ (Fig. 4.3 B). Apparently, the simple sequence structure facilitated object recognition.

The results of experiments 2, 4, and 5 are consistent with somewhat lower learning rates and reduced recognition probabilities in the range of $\gamma = 0.5$ to 0.9 (Figs. 4.3 CEF). The learning rates appear to decrease with increasing object number, with $\epsilon \approx 0.25$ in experiment 2 (8 recurring objects and 24 one-time objects), $\epsilon \approx 0.16$ in experiment 4 (10 recurring objects, 40 one-time objects), and $\epsilon \approx 0.15$ in experiment 5 (16 recurring objects, 64 one-time objects). Presumably, learning rates decrease as limited memory capacity is spread 'more thinly' over a larger number of objects.

At first glance, a second set of parameter values ($\epsilon \approx 0.5$ and $\gamma \approx 0.25$) accounts comparably well (and sometimes even better) for the experimental results (Figs. 4.3 DF). However, a closer look reveals that this 'second' fit results from an intrinsic symmetry of the model: the overall learning rate is proportional to the product of ϵ and γ and thus may be matched equally well by $(\epsilon, \gamma) \approx (0.25, 0.5)$ and by $(\epsilon, \gamma) \approx (0.5, 0.25)$. In addition, low values of γ erode the recognition probability and thus provide an indirect way of adjusting the degree of context dependence. If one introduces a further parameter to modify the relative weights of current and previous objects, comparably good fits are obtained with high values of γ (results not shown).

Finally, the results of experiment 3 are consistent with a learning rate of $\epsilon \approx 0.14$ and a wide range of recognition probabilities γ, with the best fit obtained for $\gamma \approx 0.75$.

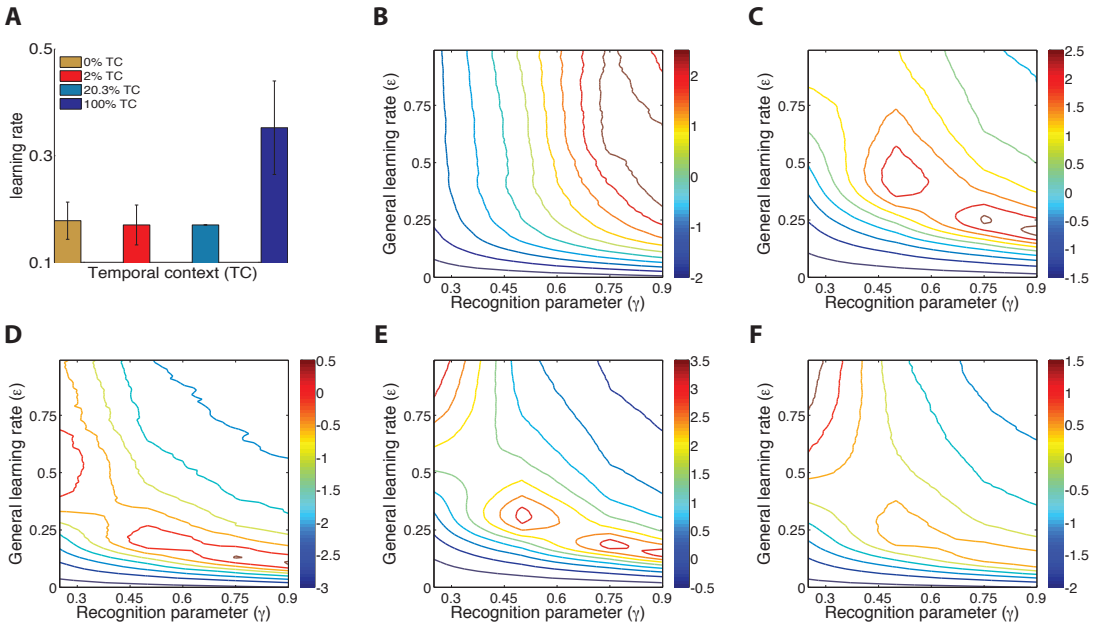

Figure 4.3: *Actual learning rates and estimated parameters.* **A:** acceleration of learning during the initial appearance of objects due to different degrees of temporal context. In the presence of a fully predictive temporal context, the accumulation of information was accelerated by 0.13 bit. Error bars show the standard deviation across experiments for each object type. Plots **B**-**F** show regions of optimal values in the parameter space (ϵ, γ), corresponding to the general learning rate and the recognition parameter, respectively. The color scales to the right of each plot refer to the fit quality f_Q for each parameter pair (ϵ, γ), which was computed as $f_Q = -\log\left(\sum_i^n (\mu_{H_i} - \mu_{M_i})^2 / (\sigma_{H_i}^2 + \sigma_{M_i}^2)\right)$, where μ_{H_i} and μ_{M_i} are the mean values of performance correct in the i-th appearance for human observers and for the model simulations, respectively, and σ_{H_i} and σ_{M_i} are the corresponding standard deviations. The higher the f_Q values, the better the fit between measured and predicted data.

The comparatively low value of ϵ reflects the memory load, which was highest in this experiment (16 recurring and 112 one-time objects).

Chapter 5

Discussion

5.1 Temporal Context Accelerates Associative Learning

We have compounded the learning of multiple visual-motor associations in various sequential orders. In every trial, the rewarded response was fully predicted by a visible visual object. Additionally, however, the rewarded response was predicted to varying degrees by the visual objects of previous trials. Five experiments showed consistently that learning is accelerated when objects of previous trials provide a predictive temporal context.

In the first experiment, the trial sequence separated object-response-pairs with and without temporal context into distinct blocks, so that the difference was evident to observers. Reaction times were significantly shorter for objects with temporal context than for objects without temporal context, indicating that observers might have applied differential cognitive strategies. In the second experiment (and all others), trials with and without temporal context were intermixed, so that the difference remained concealed from observers. Reaction time patterns showed no evidence that observers allocated attentional/memory resources differentially to trials with and without temporal context. The third experiment raised task difficulty by doubling the number

of visual objects (from 8 to 16), but confirmed the basic result: object-response-pairs with temporal context are learned faster than pairs without such context. In the fourth experiment, a partially predictive (20.3%) temporal context failed to accelerate associative learning. In the fifth and last experiment, the objects in successive trials formed ordered pairs, some predictive and others not. Only predictive pairings accelerated learning.

A number of previous studies have manipulated temporal context that (*i*) was irrelevant to the overt behavioral task and (*ii*) remained concealed from the observer. Typically, temporal context is altered by repeating a given set of trials in either fixed or random order.

In serial reaction time tasks [99], human observers respond as rapidly as possible to the locations of successive visual targets. After training, reaction times are faster when the target locations follow a repeating rather than a random pattern, which is taken as evidence of 'sequence learning' [27, 28, 116, 141]. Importantly, observers do not have to be aware of the repeating sequence in order to benefit from it [36].

In serial button press tasks [62], non-human primates are presented with pairs of visual targets and learn to press two corresponding buttons in a particular order. Both within and between daily sessions, learning is facilitated when target pairs follow each other in a repeating rather than reversed or random order [98, 115]. However, the animals do not seem to acquire choice responses for individual target pairs but rather motor sequences for 'hyper-sets' of several successive pairs [63, 115].

In visual search tasks, human observers locate a single target (which is identified by certain distinguishing characteristics) among multiple distractors. Search performance benefits from the 'spatial context' that is provided by recurring distractor configurations [24]. Interestingly, observers are unaware of the repeating configuration and contextual learning depends on an intact hippocampus [23, 25, 26]. Similar benefits accrue from the 'temporal context' created when a fixed sequence of target locations is

used in successive trials [66, 101]. This temporal effect is also implicit and appears to be mediated by visual selective attention, in that observers learn to shift attention to the next target location predicted by contextual information.

Finally, when different visual threshold discriminations (*e.g.*, contrast, motion-direction) are compounded, visual learning accelerates significantly if different displays appear in a fixed (rather than random) temporal sequence [79]. It has been proposed that predictive temporal context may facilitate the activation of an appropriate visual template for each trial [147].

The present study differed from previous investigations in a number of ways. Firstly, it forced observers to become familiar with a number of initially unfamiliar fractal patterns. This emphasis on visual recognition was modeled on paradigms developed for behaving non-human primates [91, 106, 145].

Secondly, we ensured that observers associated individual fractal patterns with particular responses and foiled alternative strategies such as acquiring motor sequences that span several successive trials. We achieved this by keeping consistent sequences short (two trials in most experiments) and by intermixing trials with different temporal contexts. This sets our situation apart from serial reaction time [99] or serial button press tasks [62].

Thirdly, observers were able to attend fully to the sole visual object presented on each trial. This stands in contradistinction to visual search paradigms, where training improves performance mainly through the anticipatory guidance of visual selective attention [66, 67, 147].

Attractor network models of associative learning [2, 64] are typically tested with electrophysiological recordings from behaving non-human primates [50, 51, 84, 133, 136]. However, behavioral observations from human observers can also furnish useful evidence, at least with respect to the more qualitative predictions of these theories. For example, behavioral experiments with sequences of self-similar images suggest that ini-

tially distinct classes of objects in associative memory become merged when exemplars of the two classes are repeatedly presented in the same temporal order [112, 137]. This confirms the qualitative prediction that events occurring consistently in the same temporal order are eventually subsumed under one and the same event class in associative memory [3, 12, 20, 57, 113].

We have presented behavioral evidence that is consistent with another qualitative prediction of attractor network models, namely, the persistent representation of past events ('delay activity'). Patterning our behavioral situation on established paradigms of conditional associative learning, we have demonstrated that the presence of consistent temporal context significantly improves choice performance. This finding implies that not just the representation of a current event, but also the representations of past events, are reinforced during conditional associative learning.

5.2 Comparison with Ideal Learner

An ideal learner is someone who has full knowledge of the structure of reward contingencies and who narrows the remaining possibilities down as quickly as possible. In our paradigm, the knowledge that each visual object deterministically predicts the rewarded action would allow an ideal learner to identify the correct action for each object after three appearances of this object.

In focusing on the current object, the ideal learner is oblivious to temporal context. While this is no disadvantage in our paradigm, it could easily develop into one in other situations. For example, consider a situation in which the current object predicts the correct action probabilistically, that is, with a probability of less than unity. In this case, it would be less than ideal to focus exclusively on the current object. Hence, an ideal learner would be open to the possibility that preceding objects may also be predictive. This kind of more 'open minded' ideal learner is realized by our model. The

downside is, of course, that this type of ideal learner is vulnerable to 'false positives', that is, to accidental configurations that are repeatedly associated with reward, without being causally predictive of the reward. An interesting extension of our work would be to compare the relative costs and benefits of considering more and more events as potentially predictive of reward.

In the presence of temporal context, human observers become more like an ideal learner only in the sense that beginning with the fourth appearance of *some* objects, observers never make a mistake associating these objects with their correct motor responses (see section 3.7). How does temporal context bring performance closer to that of an ideal learner? Does temporal context improve object recognition, for example through the anticipatory guidance of attentional and/or memory resources? Or does temporal context improve reward prediction, for example by cumulating predictive cues in the manner predicted by our model? Though interesting, these questions, unfortunately, cannot be settled on the basis of our observations. However, if the guidance of attentional/memory resources is crucial, then one would expect that improved recognition of some objects comes at the expense of other objects. In other words, when the attentional/memory load is increased (as was the case in going from experiment 2 to experiment 3), the benefit of temporal context should be diluted. Yet the behavioral results do not bear out this prediction: in both experiments, objects with temporal context (type B) enjoyed a similar advantage over objects without temporal context (types A and C). This observation suggests that the effect of temporal context does not depend on the redistribution of limited resources.

5.3 Reinforcement Learning

Our behavioral results are quantitatively consistent with a form of *model-free* reinforcement learning [33, 132]. In this approach, response choice is probabilistic, but reflects

reward expectations, which are being accumulated in the form of 'action values'. The reinforcement rule increments (decrements) these 'action values' when a chosen response receives more (less) reward than expected. The key feature is that response choice is influenced by multiple 'action values', some attaching to the object of the current trial and others attaching to objects of preceding trials (Fig. 4.2). Their effect is cumulative in the sense that the more 'action values' favor a particular response, the more likely this response is chosen. Accordingly, when successive objects appear in a consistent order, more than one 'action value' will favor the correct response, which will therefore be chosen more frequently.

The model accounts qualitatively and quantitatively for our behavioral observations, provided suitable values are chosen for learning rate ϵ and recognition parameter γ. The value of ϵ decreases as the number of fractal objects increases. The value is smaller than unity, which implies that observers concurrently acquire only a subset of stimulus-response pairings. Overall, the values of ϵ are consistent with the possibility that two to three pairings are being formed concurrently (*i.e.*, at the ideal learner rate), while the remaining pairings are being ignored. The value of γ also decreases with the number of fractal objects, consistent with growing uncertainty about object identity.

In the present series of experiments, the task set remained stable in the sense that the same stimulus-response pairings were rewarded throughout each trial sequence. However, stable task sets pose only a weak test of the model and its underlying assumptions. Far stronger tests can be devised with experimental designs that vary the task sets (*e.g.*, task reversal). To illustrate this point, we outline a hypothetical experiment with variable task set:

Consider trials $t-2$, $t-1$ and t with stimuli S_{t-2}, S_{t-1}, S_t and trial t with response R_t. While the overt task is to acquire the pairing $S_t \rightarrow R_t$, the model additionally reinforces the pairings $S_{t-2} \rightarrow R_t$ and $S_{t-1} \rightarrow R_t$ (Fig. 6.1). How will the model perform when either stimulus S_t is replaced by S'_t or response R_t is replaced by R'_t? In

the former case, two out of three pairings remain valid ($S_{t-2} \rightarrow R_t$ and $S_{t-1} \rightarrow R_t$), so that predicted performance remains above chance level. In the latter case, however, all pairings become invalid and predicted performance falls below chance level. Accordingly, this hypothetical experiment would test the model's key assumption, namely, the reinforcement of pairings between past stimuli and present response ($S_{t-2} \rightarrow R_t$ and $S_{t-1} \rightarrow R_t$).

As a preliminary test of this prediction, we have conducted a pilot study [61], in which the order of events was disrupted by replacing either an individual visual object ('object reversal'), or an individual rewarded action ('action reversal'), or both of them ('combined reversal') in an otherwise unchanged trial sequence. The results of this study are included in Appendix A (Figs. 6.2 and 6.3). Although preliminary, the results of this study suggest that all kind of reversal reduces performance to chance level. In other words, the results of this pilot study fail to bear out the predicted difference between 'action reversals' and 'object reversals'.

There are several ways in which a reinforcement model could be extended in order to accommodate the additional observations just described. For example, the number of 'action values' could be increased combinatorially, so that reinforcement applies not just to the pairing of a stimulus S_{t-i}, with $i = 0, 1, 2$, and a response R_t, but also to the triplet of past stimulus S_{t-i}, with $i = 1, 2$, current stimulus S_t, and a response R_t. This would capture the intuition that the influence of accumulated experience on the response probability is conditioned on the particular context provided by the current object S_t. If this object is missing, past experience does not apply and cannot guide the response.

Though feasible, this approach suffers from evident drawbacks. Firstly, the combinatorial increase of action values would lengthen training and slow the pace of learning. This follows from the *scaling property* of reinforcement learning, which holds that the time required to optimize behavior scales proportionately with the set size of both

environmental states and available actions [8, 14, 40, 72, 80, 81, 129]. Secondly, if performance suffers during the initial learning phase, it suffers even worse during the re-learning phase that follows a reversal. For during this phase a large number of 'action values' first have to be un-learned (as they have been rendered inappropriately by the reversal) before a large number of other 'action values' can be re-learned.

These considerations point to a fundamental problem of reinforcement framework: although context specificity is helpful to behavioral performance in stationary environments, it is detrimental to flexibility in non-stationary environments.

5.4 Models in the Attractor Framework

It is widely accepted that reinforcement mechanisms are optimal only if there is a predefined set of distinct states that are predictive of reward [29, 33, 42, 100, 132]. Thus, reinforcement models beg the question as to which events or combinations of events could potentially predict reward in a non-stationary environment. This brings us to the crucial question as to how our brain selects and creates neural representations for potentially reward-predicting events. An interesting approach to this question is the attractor framework, which postulates that the formation of such representations is based on temporal statistics of the environment. The key idea is that mental representations are realized by stable patterns of reverberating activity, which are stable steady-states ('attractors') in the neural dynamics of the network [2, 3, 50, 64].

A recent study of behavioral flexibility in reversal situations exemplifies the attractor framework [51, 120]. The authors of this study postulate two neural circuits, one for learning reward-relevant conditional associations ('associative network') and another for observing temporal contingencies ('context network'). The interaction between these two networks leads to the formation of distinct neural representations for different contexts. More specifically, the associative network comprises two populations

of excitatory neurons, which represent alternative stimulus-response associations. One population represents the stimulus-response associations appropriate for one context, whereas the other population codes the appropriate associations for another context. The two excitatory populations compete through a third, inhibitory population. As long as the reward predictions of one population are fulfilled, the currently dominant population will continue to suppress the other population, and new stimuli will be evaluated in the light of the experience encoded in the dominant population. However, when predicted rewards fail to materialize, the other population may gain ascendancy and behavior may now be governed by the experience accumulated in another, alternative context.

So how can a representation of context be formed, which can link all the stimulus-response associations that are rewarded in a particular context? The key idea is that different stimulus-response associations become linked on the basis of temporal statistics. Specifically, as long as one context holds for much longer than one trial, stimulus-response associations within this context follow each other more frequently than stimulus-response associations in different contexts. This correlational difference can be translated by Hebbian mechanisms into selective meta-associations between the stimulus-response associations of a given context. Mechanistically, the formation of these meta-associations relies on the temporal overlap between the representation of a current stimulus-response association and lingering representations of stimulus-response associations in the recent past. Further details can be found in Rigotti *et al.* [120].

Although most attempts to test the attractor framework experimentally have used single-unit recordings in behaving, non-human primates, we believe that this framework makes some predictions even at the behavioral level. For example, the neurophysiological findings of Miyashita [91] and Yakovlev *et al.* [145] imply that reverberative delay activity exists only after an attractor representation has formed. In the context of our paradigm, this suggests that lingering representations of past events are available only

after these past events have become familiar. On this basis, we would expect that the presence of consistent predecessor objects becomes influential only after these objects have become familiar and are recognized. Accordingly, it would be an interesting extension of the present study to examine whether the facilitative effect of temporal context is conditional on correct performance with regard to predecessor objects.

5.5 Cyclic Order: Another Kind of Temporal Information

In the experiments presented here, observers experienced sequences of visual objects and motor responses which exhibited different kinds of temporal structure. One type of temporal structure, which we controlled and analyzed explicitly, was temporal correlations between objects and their immediate predecessors. Another type of temporal structure was the interval between two successive appearances of the same object. This second type of temporal structure, which we did not manipulate systematically, can be termed 'cyclic order'. Note that both types of temporal structure are inter-dependent. When an object has a consistent predecessor, this implies that both object and predecessor re-appear after the same interval and thus have the same distribution of cyclic orders.

Standard associative analysis of Pavlovian conditioning has focused on the formation of a predictive relation between the conditioned stimulus (CS) and the unconditioned stimulus (US) rather than on appropriate timing [54]. In one common paradigm of Pavlovian conditioning, the onset of a tone (CS) was designed to predict a weak electric shock (US) to the skin surrounding the eye of a rabbit, which in turn caused the rabbit to make an eye blink. In addition to the well-established finding that conditioning makes the rabbit *now* blink to the tone onset, it has been observed that the time period between the tone onset and the learned blink *approximately* equals the latency between the tone onset and the shock, termed as the CS-US interval [54, 73, 74, 139, 140].

Moreover, the ratio between the CS-US interval and the interval between two successive shocks, termed as US-US interval, determines how fast the association between the tone onset and the eye blink of the rabbit forms [53]. Another common paradigm of Pavlovian conditioning establishes that a pigeon receives, *but only sometimes*, a reward (e.g. grain) through pecking an illuminated key on the wall. Hence, the pigeon keeps on pecking whether it got a reward or not. Yet, when the time between successive appearances of the reward is fixed, pigeons stop pecking immediately after the last appearance of reward and they wait approximately half the fixed interval before they begin to peck again [37].

As these examples show, animals are able to learn the time interval between two events and benefit when this time interval remains fixed. With respect to our work, the cyclic order of a visual object can be seen as time intervals. Fixing the cyclic order of a specific object provides more information as to when this object will recur. Conceivably, this regular recurrence may contribute to accelerated learning. In short, the appearance of an object is predicted both by cyclic order and by episodic context.

In our current paradigm, observers managed to learn fully predicted objects on the basis of episodic context. However, reversal situations impose additional source of uncertainty, regarding learning rates (before and after reversal) and the reliability of temporal information. An fMRI study showed that human subjects adjust their learning rates according to the volatility of the environment [9]. Which kind of temporal information is most likely to be exploited in situations with multiple 'regularly-spaced' reversals? Could observers in such scenarios learn to predict context on the basis of the cyclic order of reversals? If so, this would be a different mechanism than that suggested by the attractor framework. However, further work is needed, in order to clarify this point.

5.6 Generalizing Experience in the Reinforcement and Attractor Frameworks

Standing at the heart of RL theory, the Rescorla & Wagner [118] model postulates that learning is driven by the discrepancy between what is expected and what actually transpires. It is evident that animals (and humans) are able to generalize previous experience and to derive some expectations even for novel settings. This has been investigated in different species, for example, rats [96], birds (European starlings) [55], non-human primates [102], and humans [87].

Reinforcement models do not explain how experience can be generalized and transferred from a familiar context, where it was acquired, to an unfamiliar context, where it may neverthless prove helpful [33]. This question is one of the most difficult problems in learning theory [54, 77, 111]. The attractor framework, however, does offer at least some rudimentary account for generalization. First, the pattern-completion property (section 1.3.1) of attractor states already provides a foundation for a (very limited) degree of generalization [3, 145]. Second, and more important, the possibility of linking stimulus-response assocations into context-specific meta-associations (section 5.4) serves not only behavioral flexibility but also offers a way of activating experience acquired in one context in another, unfamiliar context [51, 120].

Chapter 6

Conclusions

Studying temporal context effects with human observers poses a number of difficulties. Humans are particularly adept at developing cognitive strategies that allocate neural resources in a task-dependent manner. In general, it cannot be assumed that human observers will apply the same neural resources to different task situations. To compare different task situations in a meaningful way, a stable allocation of resources must therefore be assured.

The present study undertook several measures to this end. Attentional allocation was stabilized by presenting only one visual object on each trial. The presence of temporal context was concealed by intermixing recurring objects with context, recurring objects without context, and one-time-objects. In addition, trial sequences were terminated before the existence of different object types could become apparent to the observer. We believe accordingly that we have developed a promising approach to studying temporal context effects with human observers.

Our results imply that not just the representation of a current event, but also the representations of past events, are reinforced during conditional associative learning. In addition, these findings are broadly consistent with the prediction of attractor network models of associative learning and their prophecy of a persistent representation of past objects.

Appendix: reversal learning

Reversal learning is a widely used approach to study context-dependent learning effects. In such a paradigm, the context is manipulated (by the experimenter), in order to discern control strategies [29]. This can be achieved either by changing the meaning of contextual cues [4] or as a result of modifications in the motivational state of the subject [5, 6, 32]. In a typical reversal learning paradigm, subjects learn to associate different stimuli with their corresponding motor responses on the basis of changing reward characteristics. The learning process comprises two phases: the learning and the re-learning phase. At the beginning, a given stimulus S evokes a reward only if associated with a specific motor response R (learning phase). However, at some point, reward contingencies change in such a way that the same stimulus S now asks for a different motor response R' (re-learning phase). Importantly, the timing and nature of reversals are not known to the observer. After the reversal, observers will receive a reward only if they unlearn the old and relearn the new associations.

In our experiments, both visual objects and rewarded motor actions followed a consistent temporal order. Accordingly, it was not possible to dissociate the relative importance of the episodic context provided by preceding objects and that provided by preceding motor actions. As a first step towards resolving this ambiguity, we conducted a pilot experiment in which the sequence of events was modified once.

As before, we created sequences in which different types of objects (type A, type B, type C) recurred 12 times each (see experiment 2 in the "Methods" chapter). One or

two objects were selected for reversal. On the 7th to 12th recurrence of these object, we either altered the rewarded response ('action reversal'), or exchanged two objects for each other or for novel objects ('object reversal'), or both ('combined reversal'). As the findings of this pilot study remain preliminary, we have not described them in full. With this qualification, we can briefly summarize the results as follows:

1. Contrary to the predictions of our reinforcement model, any type of reversal reduced performance to chance level. (Fig. 6.2).

2. The rate of recovery seemed to differ between reversal types, appearing to be faster for an 'object reversal' than for an 'action' or 'combined reversal'. (Figs. 6.2 and 6.3).

3. Compounding an 'object reversal' with a second reversal on the preceding trial also suggested that different types of temporal context are not of equal importance: the rate of recovery appeared to be faster when the preceding motor response, rather than the preceding visual object, was retained (Fig. 6.3 B).

Taken together, these preliminary findings may suggest that episodic contexts formed by both sensory and motor events facilitate associative learning, but that at least in our paradigm the dominant factor may be the context constituted by rewarded motor actions.

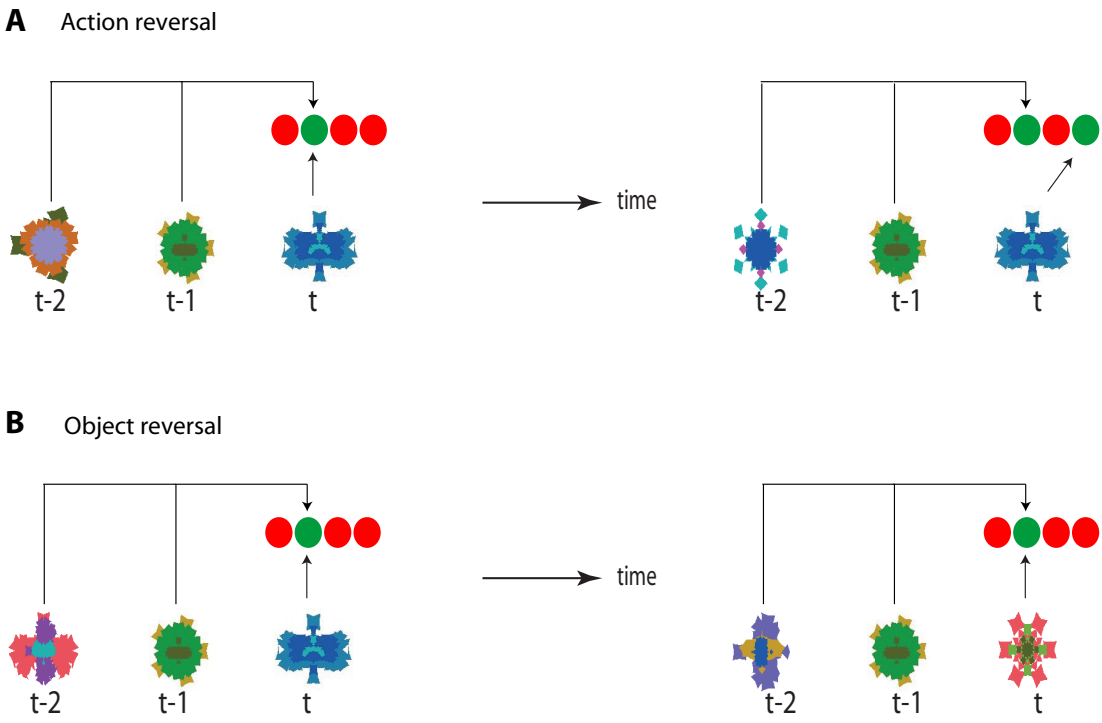

Figure 6.1: *Model's predictions for 'action' and 'object' reversals (schematic).* Let S_{t-2} and S_{t-1} be the visual stimuli presented at trials $t-2$ and $t-1$, respectively. S_t is the target stimulus at trial t with motor response R_t. **A:** learned response R_t for the target stimulus is replaced by response R'_t in the second run of the object sequence ('action reversal'). Before reversal, the model reinforces, in addition to the paring $(S_t \rightarrow R_t)$, the pairings $(S_{t-1} \rightarrow R_t)$ and $(S_{t-2} \rightarrow R_t)$. After reversal, however, these pairings become invalid, as the model has to learn the new response R'_t. Hence, the model's performance is expected to fall to chance level. **B:** target stimulus S_t is replaced by stimulus S'_t, which has the same response as that of S_t. Before reversal, the model reinforces the pairings $(S_t \rightarrow R_t)$, $(S_{t-1} \rightarrow R_t)$, and $(S_{t-2} \rightarrow R_t)$. These pairings remain valid after reversal. Hence, predicted performance remains above chance level. Simulation results are plotted in (Fig. 6.2).

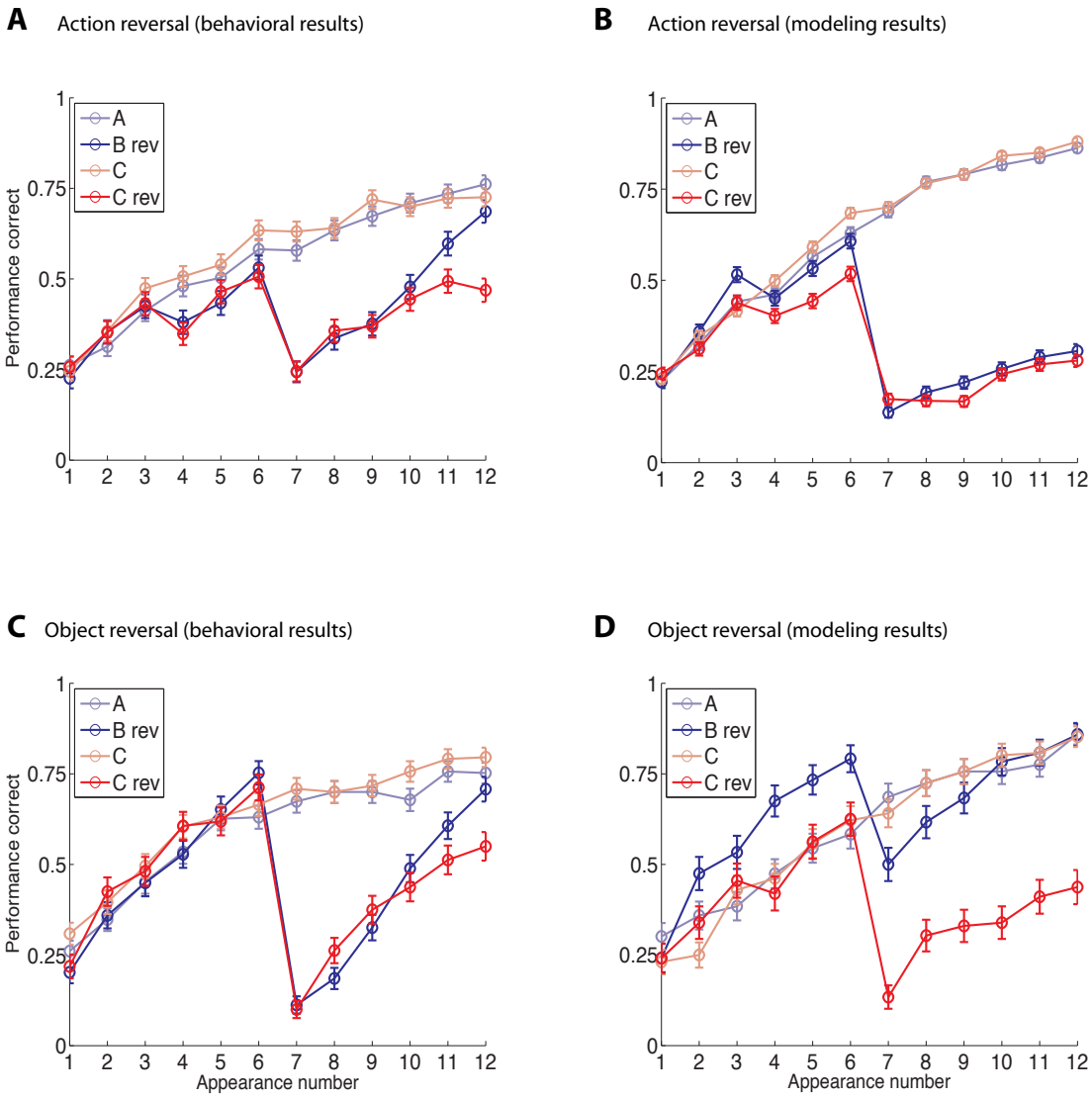

Figure 6.2: *Behavioral and modeling results for 'action' and 'object' reversals with the same objects.* Thirty two objects were used to create sequences of 144 trials. Eight of these objects were of the recurring kind and the rest were one-time objects. Each of the recurring objects appeared 12 times along the whole trial sequence (see experiment 2 in Methods and Fig. 6.1). No novel objects were included. Subplots **A** and **B** show behavioral and modeling results, respectively, for 'action reversal'. Behavioral and modeling results for 'object reversal' are plotted in **C** and **D**, respectively.

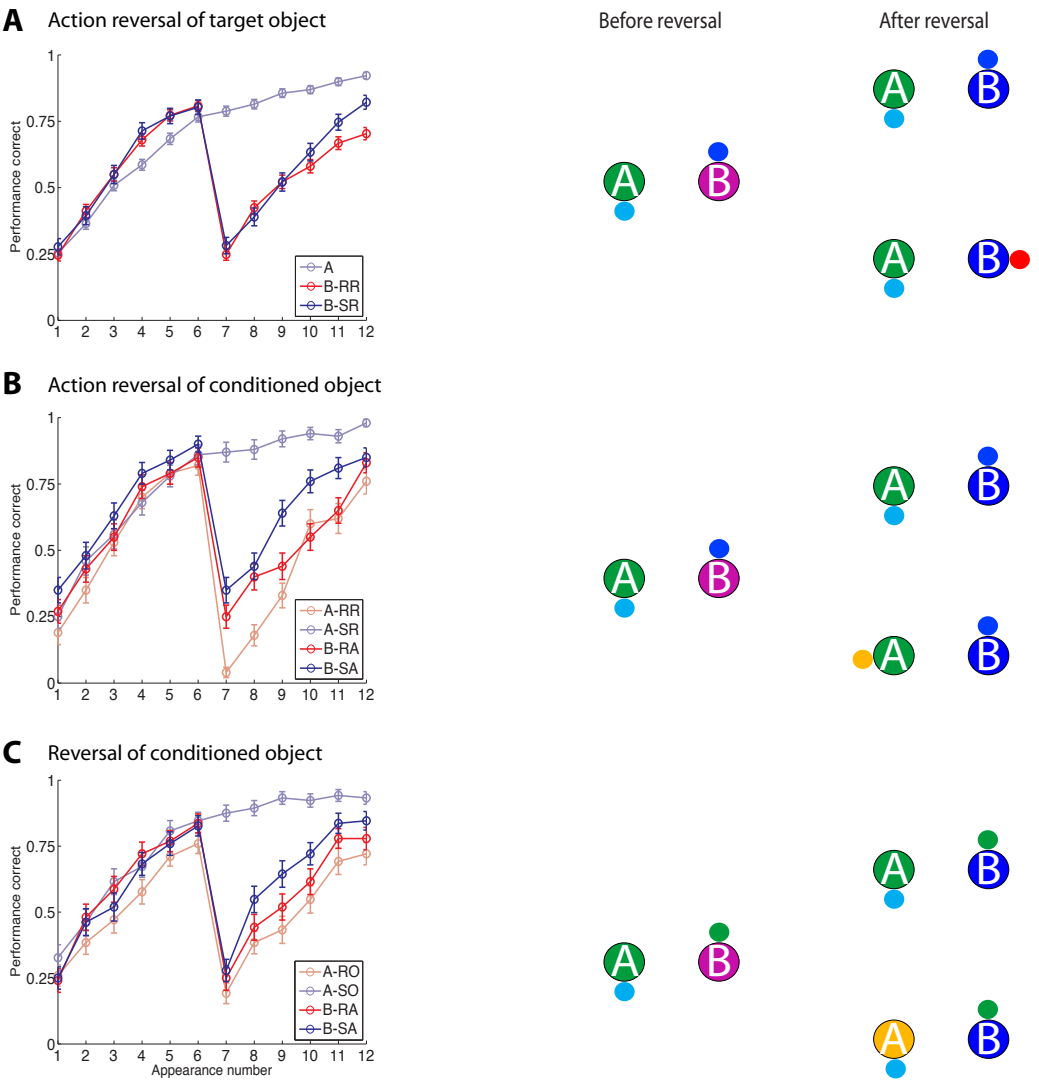

Figure 6.3: *Behavioral results for reversals with novel objects.* In all experiments, type B objects were replaced by novel objects after reversal. **A:** 'combined reversal': some of the new type B objects had the same responses as their ancestors (B-SR), whereas others not (B-RR). **B:** another 'combined reversal': response was reversed for some of the predicting type A objects. Consequently, there were four categories of objects: type A objects with reversed response (A-RR), type A with the same response (A-SR), a novel type B object whose predictor's response was reversed (B-RA), and a novel type B object whose predictor's response remained the same (B-SA). **C:** 'object reversal': some of type A objects were replaced by novel objects in the second one, while having the same response as their ancestors. As a result, there were four categories of objects: remaining type A objects (A-SO), novel type A objects (A-RO), novel type B objects with novel predictors (B-RA), and novel type B objects with their predictors unchanged (B-SA).

References

[1] Amit, D.J. (1995). The Hebbian paradigm reintegrated: local reverberations as internal representations. *Behav. Brain Sci.*, **18**, 617–626. 21, 22, 23, 33

[2] Amit, D.J., Brunel, N. & Tsodyks, M.V. (1994). Correlations of cortical hebbian reverberations: theory versus experiment. *J. Neurosci.*, **14**, 6435–6445. 14, 16, 21, 23, 33, 36, 77, 82

[3] Amit, D.J., Fusi, S. & Yakovlev, V. (1997). Paradigmatic working memory (attractor) cell in it cortex. *Neural Comput.*, **9**, 1071–1092. 14, 18, 21, 22, 23, 33, 78, 82, 86

[4] Asaad, W.F., Rainer, G. & Miller, E.K. (1998). Neural activity in the primate prefrontal cortex during associative learning. *Neuron*, **21**, 1399–1407. 13, 16, 26, 88

[5] Balleine, B.W. & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacol.*, **37**, 407 – 419. 13, 26, 88

[6] Balleine, B.W., Delgado, M.R. & Hikosaka, O. (2007). The role of the dorsal striatum in reward and decision-making. *J. Neurosci.*, **27**, 8161–8165. 26, 28, 88

[7] Barbieri, F. & Brunel, N. (2008). Can attractor network models account for the statistics of firing rates during persistent activity in prefrontal cortex? *Front. Neurosci.*, **2**, 114–122. 21, 33

[8] Barto, A.G. & Mahadevan, S. (2003). Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamic Systems: Theory and Applications*, **13**, 343–379. 16, 82

[9] Behrens, T.E.J., Woolrich, M.W., Walton, M.E. & Rushworth, M.F.S. (2007). Learning the value of information in an uncertain world. *Nat. Neurosci.*, **10**, 1214–1221. 69, 85

[10] Belova, M.A., Paton, J.J. & Salzman, C.D. (2008). Moment-to-moment tracking of state value in the amygdala. *J. Neurosci.*, **28**, 10023–10030. 13

[11] Bertsekas, D.P. & Tsitsiklis, J.N. (1996). *Neuro-Dynamic Programming*. Athena Scientific. 12, 25, 31

[12] Blumenfeld, B., Preminger, S. & Sagi, D. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, **52**, 38 394. 20, 78

[13] Boettiger, C.A. & D'Esposito, M. (2005). Frontal networks for learning and executing arbitrary stimulus-response associations. *J. Neurosci.*, **25**, 2723–2732. 16

[14] Botvinick, M.M., Niv, Y. & Barto, A.C. (2009). Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective. *Cognition*, **113**, 262 – 280. 16, 82

[15] Boussaoud, D. & Wise, S.P. (1993). Primate frontal cortex: effects of stimulus and movement. *Exp. Brain Res.*, **95**, 28–40. 13

[16] Brasted, P.J. & Wise, S.P. (2004). Comparison of learning-related neuronal activity in the dorsal premotor cortex and striatum. *Eur. J. Neurosci.*, **19**, 721–740. 16

[17] Brasted, P.J., Bussy, T.J., Murray, E.A. & Wise, S.P. (2003). Role of the hippocampal system in associative learning beyond the spatial domain. *Brain*, **126**, 1202–1223. 16

[18] Braun, J. & Mattia, M. (2010). Attractors and noise: Twin drivers of decisions and multistability. *NeuroImage*, **52**, 740 – 751, computational Models of the Brain. 23

[19] Brovelli, A., Laksiri, N., Nazarian, B., Meunier, M. & Boussaoud, D. (2008). Understanding the neural computations of arbitrary visuomotor learning through fmri and associative learning theory. *Cereb. Cortex*, **18**, 1485–1495. 16

[20] Brunel, N. (1996). Hebbian learning of context in recurrent neural networks. *Neural Comput.*, **8**, 1677–1710. 21, 22, 23, 33, 78

[21] Bunge, S.A., Wallis, J.D., Parker, A., Brass, M., Crone, E.A., Hoshi, E. & Sakai, K. (2005). Neural circuitry underlying rule use in humans and nonhuman primates. *J. Neurosci.*, **25**, 10347–10350. 16

[22] Bush, R.R. & Mosteller, F. (1951). A mathematical model for simple learning. *Psychol. Rev.*, **58**, 313–323. 29

[23] Chun, M.M. (2000). Contextual cueing of visual attention. *Trends Cognit. Sci.*, **4**, 170–178. 76

[24] Chun, M.M. & Jiang, Y. (1998). Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognit. Psychol.*, **36**, 28–71. 20, 76

[25] Chun, M.M. & Jiang, Y. (2003). Implicit, long-term spatial contextual memory. *J. Exp. Psychol. Learn. Mem. Cognit.*, **29**, 224–234. 20, 76

[26] Chun, M.M. & Phelps, E.A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nat. Neurosci.*, **2**, 844–847. 76

[27] Cohen, A., Ivry, R. & Keele, S. (1990). Attention and structure in sequence learning. *J. Exp. Psychol. Learn. Mem. Cognit.*, **16**, 17–30. 20, 76

[28] Curran, T. & Keele, S.W. (1993). Attentional and nonattentional forms of sequence learning. *J. Exp. Psychol. Learn. Mem. Cognit.*, **19**, 189–202. 20, 76

[29] Daw, N., Niv, Y. & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.*, **8**, 1704–1711. 13, 17, 24, 25, 26, 28, 82, 88

[30] Dayan, P. (2008). The role of value systems in decision making. In C. Engel & W. Singer, eds., *Better than Conscious? Decision Making, the Human Mind, and Implications for Institutions*, 50–71, The MIT Press, Frankfurt, Germany. 27

[31] Dayan, P. & Abbott, L.F. (2005). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press. 24, 25, 30, 50

[32] Dayan, P. & Balleine, B.W. (2002). Reward, motivation, and reinforcement learning. *Neuron*, **36**, 285–298. 26, 28, 88

[33] Dayan, P. & Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Curr. Opin. Neurobiol.*. 24, 26, 27, 79, 82, 86

[34] Dayan, P., Kakade, S. & Montague, P.R. (2000). Learning and selective attention. *Nat. Neurosci.*, **3**, 1218–1223. 30, 69, 71

[35] Desimone, R., Albright, T.D., Gross, C.G. & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.*, **4**, 2051–2062. 16

[36] Destrebecqz, A. & Cleeremans, A. (2001). Can sequence learning be implicit? new evidence with the process dissociation procedure. *Psychon. Bull. Rev.*, **8**, 343–350. 76

[37] Dews, P.B. (1970). The theory of fixed-interval responding. In W.N. Schoenfeld, ed., *The Theory of Reinforcement Schedules*, 43–61, Appleton-Century-Crofts, New York: Appleton-Century-Crofts. 85

[38] Dickinson, A. (1980). *Contemporary Animal Learning Theory*. Cambridge Univ. Press, Cambridge. 13

[39] Dickinson, A. (1994). Instrumental conditioning. In N.J. Mackintosh, ed., *Animal Learning and Cognition*, 45–80, Academic Press, London. 29

[40] Dietterich, T.G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *J. Artif. Intell. Res.*, **13**, 227–303. 16, 82

[41] Doya, K. (2002). Metalearning and neuromodulation. *Neural Netw.*, **15**, 495–506. 69

[42] Doya, K. (2007). Reinforcement learning: Computational theory and biological mechanisms. *HFSP J.*, **1**, 30–40. 24, 82

[43] Eacott, M.J. & Gaffan, D. (1992). Inferotemporal-frontal disconnection: the uncinate fascicle and visual associative learning in monkeys. *Eur. J. Neurosci.*, **4**, 1320–1332. 16

[44] Eichenbaum, H., Yonelinas, A.P. & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annu. Rev. Neurosci.*, **30**, 123–152. 16

[45] Eliassen, J.C., Souza, T. & Sanes, J.N. (2003). Experience-dependent activation patterns in human brain during visual-motor associative learning. *J. Neurosci.*, **23**, 10540–10547. 16

[46] Fanselow, M.S. (1986). Associative vs topographical accounts of the immediate shock-freezing deficit in rats: Implications for the response selection rules governing species-specific defensive reactions. *Learning and Motivation*, **17**, 16 – 39. 14

[47] Fanselow, M.S. (2007). Context: What's so special about it? In H.L.R. III, Y. Dudai & S.M. Fitzpatrick, eds., *Science of Memory: Concepts*, 101–105, Oxford University Press, Inc., New York. 13

[48] Franklin, K. & McCoy, S. (1979). Pimozide-induced extinction in rats: Stimulus control of responding rules out motor deficit. *Pharmacology Biochemistry and Behavior*, **11**, 71–75. 32

[49] Freedman, D.J., Riesenhuber, M., Poggio, T. & Miller, E.K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *J. Neurosci.*, **23**, 5235–5246. 16

[50] Fusi, S., Drew, P.J. & Abbott, L.F. (2005). Cascade models of synaptically stored memories. *Neuron*, **45**, 599–611. 77, 82

[51] Fusi, S., Asaad, W., Miller, E. & Wang, X.J. (2007). A neural circuit model of flexible sensorimotor mapping: learning and forgetting on multiple timescales. *Neuron*, **54**, 319–333. 33, 77, 82, 86

[52] Gaffan, D. & Harrison, S. (1988). Inferotemporal-frontal disconnection and fornix transection in visuomotor conditional learning by monkeys. *Behav. Brain Res.*, **31**, 149–163. 15

[53] Gallistel, C.R. & Gibbon, J. (2000). Time, rate, and conditioning. *Psychol. Rev.*, **107**, 289–344. 85

[54] Gallistel, C.R. & King, A.P. (2009). *Memory and the Computational Brain*. Wiley-Blackwell, West Sussex, United Kingdom, 1st edn. 84, 86

[55] Gentner, T.Q., Fenn, K.M., Margoliash, D. & Nusbauml, H.C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, **440**, 1204–1207. 86

[56] Gordon, W.C. & Klein, R.L. (1994). Animal memory: The effects of context change on retention performance. In N.J. Mackintosh, ed., *Animal Learning and Cognition*, 255–280, Academic Press, London. 13

[57] Griniasty, M., Tsodyks, M.V. & Amit, D.J. (1993). Conversion of temporal correlations between stimuli to spatial correlations between attractors. *Neural Comput.*, **5**, 1–17. 21, 78

[58] Hadj-Bouziane, F., Meunier, M. & Boussaoud, D. (2003). Conditional visuo-motor learning in primates: a key role for the basal ganglia. *J. Physiol. Paris*, **97**, 567–579. 16

[59] Hadj-Bouziane, F., Frankowska, H., Meunier, M., Coquelin, P. & Boussaoud, D. (2006). Conditional visuo-motor learning and dimension reduction. *Cogn. Process.*, **7**, 95–104. 16

[60] Hall, G. (1994). Pavlovian conditioning: Laws of association. In N.J. Mackintosh, ed., *Animal Learning and Cognition*, 15–44, Academic Press, London. 14

[61] Hamid, O.H. & Braun, J. (2010). Relative importance of sensory and motor events in reinforcement learning. In *Perception* **39**, ECVP Abstract Supplement, page 48. 81

[62] Hikosaka, O., Rand, M.K., Miyachi, S. & Miyashita, K. (1995). Learning of sequential movements in the monkey: process of learning and retention of memory. *J. Neurophysiol.*, **74**, 1652–1661. 20, 76, 77

[63] Hikosaka, O., Nakamura, K., Sakai, K. & Nakahara, H. (2002). Central mechanisms of motor skill learning. *Curr. Opin. Neurobiol.*, **12**, 217–222. 20, 76

[64] Hopfield, J.J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, **79**, 2554–2558. 16, 77, 82

[65] Hubert, M. & Kenning, P. (2009). Im Kopf des Konsumenten. *Gehirn und Geist*, **1**, 44–49. 17

[66] Jiang, Y. & Chun, M.M. (2001). Selective attention modulates implicit learning. *Quart. J. Exp. Psychol.*, **54**, 1105–1124. 77

[67] Jiang, Y. & Leung, A.W. (2005). Implicit learning of ignored visual context. *Psychon. Bull. Rev.*, **12**, 100–106. 77

[68] Kacelnik, A. (1998). Normative and descriptive models of decision making: time discounting and risk sensitivity. In M. Oaksford & N. Chater, eds., *Rational Models of Cognition*, 54–70, Oxford University Press, Inc., Oxford, New York, Tokyo. 24

[69] Kaelbling, L.P., Littman, M.L. & Moore, A.W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, **4**, 237–285. 12, 16, 24, 26, 27

[70] Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME Journal of Basic Engineering*, 35–45. 69

[71] Kamin, L.J. (1969). Predictability, surprise, attention, and conditioning. In B.A. Campbell & R.M. Church, eds., *Punishment and Aversive Behavior*, 242–259, Appleton-Century-Crofts, New York. 29, 30

[72] Kearns, M. & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, **49**, 209–232. 16, 82

[73] Kehoe, E.J., Graham-Clarke, P. & Schreurs, B.G. (1989). Temporal patterns of the rabbit's nictitating membrane response to compound and component stimuli under mixed cs-us intervals. *Behavioral Neuroscience*, **103**, 283–295. 84

[74] Kehoe, E.J., Ludvig, E.A. & Sutton, R.S. (2010). Timing in trace conditioning of the nictitating membrane response of the rabbit (oryctolagus cuniculus): Scalar, nonscalar, and adaptive features. *Learn. Mem.*, **17**, 600–604. 84

[75] Kiernan, M., Westbrook, R. & Cranney, J. (1995). Immediate shock, passive avoidance, and potentiated startle: Implications for the unconditioned response to shock. *Animal Learning and Behavior*, **23**, 22–30. 14

[76] Killcross, S. & Coutureau, E. (2003). Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex*, **13**, 400–408. 13

[77] Koechlin, E. & Hyafil, A. (2007). Anterior prefrontal function and the limits of human decision-making. *Science*, **318**, 594–598. 86

[78] Kremer, E.F. (1978). The Rescorla-Wagner model: losses in associative strength in compound conditioned stimuli. *J. Exp. Psychol. Animal Behav. Proc.*, **4**, 22–36. 30

[79] Kuai, S.G., Zhang, J.Y., Klein, S.A., Levi, D.M. & Yu, C. (2005). The essential role of stimulus temporal patterning in enabling perceptual learning. *Nat. Neurosci.*, **8**, 1497–1499. 77

[80] Li, L. & Littman, M.L. (2010). Reducing reinforcement learning to KWIK online regression. *Ann. Math. Artif. Intell.*, **58**, 217–237. 16, 82

[81] Li, L., Walsh, T.J. & Littman, M.L. (2006). Towards a unified theory of state abstraction for mdps. In *Ninth International Symposium on Artificial Intelligence and Mathematics*, Florida Atlantic University. 16, 82

[82] Ljungberg, T., Apicella, P. & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *J. Neurophysiol.*, **67**, 145–163. 33

[83] Logothetis, N.K., Pauls, J. & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, **5**, 552–563. 16

[84] Loh, M. & Deco, G. (2005). Cognitive flexibility and decision-making in a model of conditional visuomotor associations. *Eur. J. Neurosci.*, **22**, 2927–2936. 77

[85] Mackintosh, N.J. (1983). *Conditioning and Associative Learning*. Oxford Univ. Press, Oxford. 13

[86] Maia, T.V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cogn. Affect. Behav. Neurosci.*, **9**, 343–64. 27

[87] Marcus, G.F., Vijayan, S., Rao, S.B. & Vishtonl, P.M. (1999). Rule learning by seven-month-old infants. *Science*, **283**, 77–80. 86

[88] Marr, D. (2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press, Cambridge, Massachusetts, London, England. 23

[89] Mars, P., Chen, J.R. & Nambiar, R. (1996). *Learning Algorithms: Theory and Applications in Signal Processing, Control and Communications*. CRC Press, Inc. 12

[90] Miller, E.K., Freedman, D.J. & Wallis, J.D. (2002). The prefrontal cortex: categories, concepts and cognition. *Phil. Trans. R. Soc. Lond. B Biol. Sci.*, **357**, 1123–1136. 13, 16

[91] Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, **335**, 817–820. 14, 18, 19, 20, 21, 33, 35, 36, 38, 52, 77, 83

[92] Miyashita, Y. & Chang, H.S. (1988). Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, **331**, 68–70. 18, 19, 33

[93] Montague, P.R., Dayan, P. & Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.*, **16**, 1936–1947. 33

[94] Morrison, S.E. & Salzman, C.D. (2009). The convergence of information about rewarding and aversive stimuli in single neurons. *J. Neurosci.*, **29**, 11471–11483. 13

[95] Muhammad, R., Wallis, J.D. & Miller, E.K. (2006). A comparison of abstract rules in the prefrontal cortex, premotor cortex, the inferior temporal cortex and the striatum. *J. Cognit. Neurosci.*, **18**, 974–989. 13, 16

[96] Murphy, R.A., Mondragn, E. & Murphyl, V.A. (2008). Rule learning by rats. *Science*, **319**, 1849–1851. 86

[97] Murray, E.A., Bussey, T.J. & Wise, S.P. (2000). Role of prefrontal cortex in a network for arbitrary visuomotor mapping. *Exp. Brain Res.*, **133**, 114–129. 13, 16

[98] Nakahara, H., Doya, K. & Hikosaka, O. (2001). Parallel cortico-basal ganglia mechanisms for acquisition and execution of visuomotor sequences: a computational approach. *J. Cognit. Neurosci.*, **13**, 626–647. 76

[99] Nissen, M.J. & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognit. Psychol.*, **19**, 1–32. 76, 77

[100] Niv, Y. & Montague, P.R. (2008). Theoretical and empirical studies of learning. In P.W. Glimcher, C. Camerer, E. Fehr & R. Poldrack, eds., *Neuroeconomics: Decision Making and The Brain*, 329–349, NY: Academic Press, New York. 24, 25, 30, 31, 33, 34, 82

[101] Olson, I.R. & Chun, M.M. (2001). Temporal contextual cuing of visual attention. *J. Exp. Psychol. Learn. Mem. Cognit.*, **27**, 1299–1313. 77

[102] Orlov, T., Yakovlev, V., Hochstein, S. & Zohary, E. (2000). Macaque monkeys categorize images by their ordinal number. *Nature*, **404**, 77–80. 86

[103] Owen, A.M. (1997). Cognitive planning in humans: neuropsychological, neuroanatomical and neuropharmacological perspectives. *Prog. Neurobiol.*, **53**, 431–450. 28

[104] Packard, M.G. & Knowlton, B. (2002). Learning and memory functions of the basal ganglia. *Ann. Rev. Neurosci.*, **25**, 563–593. 28

[105] Parris, B.A., Thai, N.J., Benattayallah, A., Summers, I.R. & Hodgson, T.L. (2007). The role of the lateral prefrontal cortex and anterior cingulate in stimulus-response association reversals. *J. Cognit. Neurosci.*, **19**, 13–24. 16

[106] Pasupathy, A. & Miller, E.K. (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature*, **433**, 873–876. 16, 77

[107] Paton, J.J., Belova, M.A., Morrison, S.E. & Salzman, C.D. (2006). The primate amygdala represents the positive and negative value of visual stimuli during learning. *Nature*, **439**, 865–870. 13

[108] Pearce, J.M. & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.*, **87**, 532–552. 29

[109] Petrides, M. (1987). Conditional learning and the primate frontal cortex. In E. Perecman, ed., *The Frontal Lobes Revisited*, 91–108, The IRBN Press, New York. 15

[110] Pfeifer, R. & Scheier, C. (2001). *Understanding Intelligence*. The MIT Press, Cambridge, Massachusetts, London, England. 12

[111] Phelps, E.A. (2007). Learning: Challenges in the merging of levels. In H.L.R. III, Y. Dudai & S.M. Fitzpatrick, eds., *Science of Memory: Concepts*, 45–48, Oxford University Press, Inc., New York. 86

[112] Preminger, S., Sagi, D. & Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects. *Vision Res.*, **47**, 965–973. 20, 35, 78

[113] Preminger, S., Blumenfeld, B., Sagi, D. & Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proc. Natl. Acad. Sci. USA*, **106**, 5371–5376. 20, 78

[114] Rainer, G., Rao, S.C. & Miller, E.K. (1999). Prospective coding for objects in primate prefrontal cortex. *J. Neurosci.*, **19**, 5493–5505. 20

[115] Rand, M.K., Hikosaka, O., Miyachi, S., Lu, X., Nakamura, K., Kitaguchi, K. & Shimo, Y. (2000). Characteristics of sequential movements during early learning period in monkeys. *Exp. Brain Res.*, **131**, 293–304. 76

[116] Reed, J. & Johnson, P. (1994). Assessing implicit learning with indirect tests: determining what is learned about sequence structure. *J. Exp. Psychol. Learn. Mem. Cognit.*, **20**, 585–594. 20, 76

[117] Rescorla, R.A. & Lolordo, V.M. (1968). Inhibition of avoidance behavior. *J. Comp. Physiol. Psychol.*, **59**, 406–412. 30

[118] Rescorla, R.A. & Wagner, A.R. (1972). A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black & W.F. Prokasy, eds., *Classical Conditioning II: Current Research and Theory*, 64–99, Appleton-Century-Crofts, New York. 29, 66, 86

[119] Reynolds, G.S. (1961). Attention in the pigeon. *J. Exp. Anal. Behav.*, **4**, 203–208. 30

[120] Rigotti, M., Rubin, D.B.D., Morrison, S.E., Salzman, C.D. & Fusi, S. (2010). Attractor concretion as a mechanism for the formation of context representations. *NeuroImage*, **52**, 833 – 847, computational Models of the Brain. 82, 83, 86

[121] Sakai, K. & Miyashita, Y. (1991). Neural organization for the long-term memory of paired associates. *Nature*, **354**, 152–155. 20, 33

[122] Sakai, K., Naya, Y. & Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learn. Mem.*, **1**, 83–105. 20, 33

[123] Salinas, E. (2004). Context-dependent selection of visuomotor maps. *BMC Neuroscience*, **5**, 47. 13, 16

[124] Schmajuk, N. & Holland, P.C. (1998). *Occasion Setting: Associative Learning and cognition in animals*. American Psychological Association. 13

[125] Schultz, W. & Romo, R. (1990). Dopamine neurons of the monkey midbrain: contingencies of responses to stimuli eliciting immediate behavioral reactions. *J. Neurophysiol.*, **63**, 607–624. 33

[126] Schultz, W., Dayan, P. & Montague, P.R. (1997). A neural substrate of prediction and reward. *Science*, **275**, 1593–1599. 30, 33, 34

[127] Shannon, C.E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423, 623–656, July, October. 44

[128] Sigala, N. & Logothetis, N.K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, **415**, 318–320. 16

[129] Sutton, R., Precup, D. & Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, **112**, 181–211. 16, 82

[130] Sutton, R.S. (1992). Gain adaptation beats least squares? In *Proceedings of the Seventh Yale Workshop on Adaptive and Learning Systems*, 161–166, Yale University, New Haven, CT. 69, 70, 72

[131] Sutton, R.S. & Barto, A.G. (1990). Time derivative models of pavlovian reinforcement. In M. Gabriel & J. Moore, eds., *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, 539–602, The MIT Press, Cambridge, MA. 30, 32

[132] Sutton, R.S. & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, Massachusetts. 24, 25, 26, 31, 35, 79, 82

[133] Szabo, M., Deco, G., Fusi, S., Giudice, P.D., Mattia, M. & Stetter, M. (2006). Learning to attend: modeling the shaping of selectivity in infero-temporal cortex in a categorization task. *Biol. Cybern.*, **94**, 351–365. 77

[134] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.*, **19**, 109–139. 16

[135] Thorndike, E.L. (1898). Animal intelligence: An experimental study of the associative processes in animals. In J.M. Cattell & J.M. Baldwin, eds., *The Psychological Review*, vol. 2, The Macmillan company, New York, London. 15, 16

[136] Vasilaki, E., Fusi, S., Wang, X.J. & Senn, W. (2009). Learning flexible sensori-motor mappings in a complex network. *Biol. Cybern.*, **100**, 147–158. 77

[137] Wallis, G. & Bulthoff, H.H. (2001). Effects of temporal association on recognition memory. *Proc. Natl. Acad. Sci. USA*, **98**, 4800–4804. 20, 35, 78

[138] Wallis, J.D. & Miller, E.K. (2003). From rule to response: neuronal processes in the premotor and prefrontal cortex. *J. Neurophysiol.*, **90**, 1790–1806. 13, 16

[139] Weidemann, G., Georgilas, A. & Kehoe, E.J. (1999). Temporal specificity in patterning of the rabbit nictitating membrane response. *Animal Learning and Behavior*, **27**, 99–107. 84

[140] White, N.E., Kehoe, E.J., Choi, J.S. & Moore, J.W. (2000). Coefficients of variation in timing of the classically conditioned eyeblink in rabbits. *Psychobiology*, **28**, 520–524. 84

[141] Willingham, D.B., Nissen, M.J. & Bullemer, P. (1989). On the development of procedural knowledge. *J. Exp. Psychol. Learn. Mem. Cognit.*, **15**, 1047–1060. 76

[142] Wirth, S., Yanike, M., Frank, L.M., Smith, A.C., Brown, E.N. & Suzuki, W.A. (2003). Single neurons in the monkey hippocampus and learning of new associations. *Science*, **300**, 1578–1581. 16

[143] Wise, R.A., Spindler, J., deWit, H., & Gerberg, G.J. (1978). Neuroleptic-induced "anhedonia" in rats: pimozide blocks reward quality of food. *Science*, **201**, 262–264. 32, 33

[144] Wise, S.P. & Murray, E.A. (2000). Arbitrary associations between antecedents and actions. *Trends Neurosci.*, **23**, 271–276. 13, 16

[145] Yakovlev, V., Fusi, S., Berman, E. & Zohary, E. (1998). Inter-trial neuronal activity in inferior temporal cortex: a putative vehicle to generate long-term visual associations. *Nat. Neurosci.*, **1**, 310–317. 18, 22, 23, 33, 35, 77, 83, 86

[146] Yanike, M., Wirth, S., Smith, A.C., Brown, E.N. & Suzuki, W.A. (2009). Comparison of associative learning-related signals in the macaque perirhinal cortex and hippocampus. *Cereb. Cortex*, **19**, 1064–1078. 16

[147] Zhang, J.Y., Kuai, S.G., Xiao, L.Q., Klein, S.A., Levi, D.M. & Yu, C. (2008). Stimulus coding rules for perceptual learning. *PLoS Biol.*, **6**, 1651–1660. 77

Cirriculum Vitae

Personal

Name	Oussama Hussein Hamid
Address	Am Fuchsberg 2, 39112 Magdeburg
Date of birth	October 27th, 1968
Place of birth	Kuwait
Marital status	Married
Nationality	German

Academic Education

01/2006 - 01/2011	Doctoral student, Department of Cognitive Biology, Faculty of Natural Sciences, Otto-von-Guericke University Magdeburg.
10/1997 - 12/2002	Studies of Computational Visualistics (Dipl.-Ing.), Faculty of Computer Science, Otto-von-Guericke University Magdeburg.
10/1995 - 09/1998	Studies of Mathematics (Vordiplom), Faculty of Mathematics, Otto-von-Guericke University Magdeburg.

Professional Experience

03/2005 - 12/2005	Research associate, Institute of Biometrics and Medical Informatics, Faculty of Medicine, University Clinic of Magdeburg.
01/2003 - 04/2005	Research associate, Clinic of Radiotherapy, Faculty of Medicine University Clinic of Magdeburg.
10/1998 - 12/2002	Student assistant, Institute for Analysis and Numerical Mathematics, Otto-von-Guericke University Magdeburg.

Honors and Awards

04/2001 - 09/2001	Internship at the School of Computer Science, Carleton University, Ottawa, Canada. Funded by NSER group.

List of Publications

Hamid O. H., Wendemuth A., Braun J. **(2010)**."Temporal context and conditional associative learning". *BMC Neuroscience 2010*, **10**:45 (1-15). DOI: 10.1186/1471-2202-11-45.

Glüge S., **Hamid O. H.**, Wendemuth A. **(2010)**. "A Simple Recurrent Network for Implicit Learning of Temporal Sequences". *Cognitive Computation*, **2**:265-271. DOI: 10.1007/s12559-010-9066-z.

Selbständigkeitserklärung

Hiermit erkläre ich, dass ich die von mir eingereichte Dissertation mit dem Thema

On the Role of Temporal Context in Human Reinforcement Learning

selbständig verfasst, nicht schon als Dissertation verwendet habe und die benutzten Hilfsmittel und Quellen vollständig angegeben habe.

Weiterhin erkläre ich, dass ich weder diese noch eine andere Arbeit zur Erlangung des akademischen Grades doctor rerum naturalium (Dr. rer. nat.) an anderen Einrichtungen eingereicht habe.

Magdeburg, den 25.01.2011