1998

Reinforcement learning in intelligent control : a biologically-inspired approach to the relearning problem

D'Cruz, Brendan

Reinforcement Learning in Intelligent Control: A Biologically-Inspired Approach to the Relearning Problem

by

Brendan D'Cruz

A thesis submitted to the University of Plymouth

in partial fulfilment for the degree of

Doctor of Philosophy

School of Computing

Faculty of Technology

May 1998

Reinforcement Learning in Intelligent Control:

A Biologically-Inspired Approach to the Relearning Problem

Brendan D'Cruz

Abstract

The increasingly complex demands placed on control systems have resulted in a need for intelligent control, an approach that attempts to meet these demands by emulating the capabilities found in biological systems. The need to exploit existing knowledge is a desirable feature of any intelligent control system, and this leads to the relearning problem. The problem arises when a control system is required to effectively learn new knowledge whilst exploiting still useful knowledge from past experiences. This thesis describes the adaptive critic system using reinforcement learning, a computational framework that can effectively address many of the demands in intelligent control, but is less effective when it comes to addressing the relearning problem. The thesis argues that biological mechanisms of reinforcement learning (and relearning) may provide inspiration for developing artificial intelligent control mechanisms that can better address the relearning problem. A conceptual model of biological reinforcement learning and relearning is presented, and the thesis shows how inspiration derived from this model can be used to modify the adaptive critic. The performance of the modified adaptive critic system on the relearning problem is investigated based on simulations of the pole balancing problem, and this is compared to the performance of the original adaptive critic system. The thesis presents an analysis of the results from these simulations, and discusses the significance of these results in terms of addressing the relearning problem.

List of Contents

Contents **Page**

List of Tables

Table **Page**

List of Figures

Figure **Page**

Acknowledgments

The author would like to express his gratitude to the following people :-

- Professor Mike Denham … for his invaluable support and considerable influence throughout this research and preparation of the thesis
- Dr. Guido Bugmann … for his relentless pragmatism and role as devil's advocate, sometimes mistakenly perceived as more devil than advocate !
- Dr. Raju Bapi … whose positive impact on the motivation and working practices of the author can least be described as "immense"
- Professor Jack Boitano … for his entertaining neurophysiology lectures at a time when they were very much needed
- Professor Andrew Barto … for his useful comments on various aspects of this work.

The existence of this thesis owes much to Professor Edvard Petrov, who endeavoured to convince the author that postgraduate academic life would be a worthwhile and rewarding experience, and not just a place to hide.

Author's Declaration

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. This study was funded with the aid of a studentship from the Engineering and Physical Sciences Research Council (EPSRC).

The author attended an EPSRC funded pre-conference workshop and subsequent conference *IEE Control'94* at the University of Warwick, March 1994.

The author attended a course of 25 lectures on "*Fundamentals of Neuroscience*" given by Professor Jack Boitano during his sabbatical with the Neural and Adaptive Systems Group at the University of Plymouth.

Publications

Bapi, R.S.; D'Cruz, B.; Bugmann, G.; (1995), "*Neuro-Resistive Grid Approach to Pole-Balancing Problem*", Proc. ICANN'95, Paris, vol.2, pp.539-544.

Bapi, R.S.; D'Cruz, B.; Bugmann, G.; (1997), "*Neuro-Resistive Grid Approach to Trainable Controllers: A Pole Balancing Example*", Neural Computing & Applications, vol.5, pp.33-44.

Signed: .....

Date: .....

Chapter One

Introduction

1.1 - Background

A general definition of a control system is a system that maintains some physical quantities more or less accurately around prescribed values [Narendra, 1994]. The physical quantities are part of a dynamical system usually referred to as the 'plant'. The plant is situated in a particular 'environment', which affects the dynamic processes in the plant by providing unpredictable disturbances to the plant. The control inputs to the plant are produced by a device called the 'controller', which is an intrinsic part of the control system. This device observes the outputs from the plant, and then modifies its own inputs to the plant to achieve the desired behaviour. Figure 1.1 illustrates this control scheme.

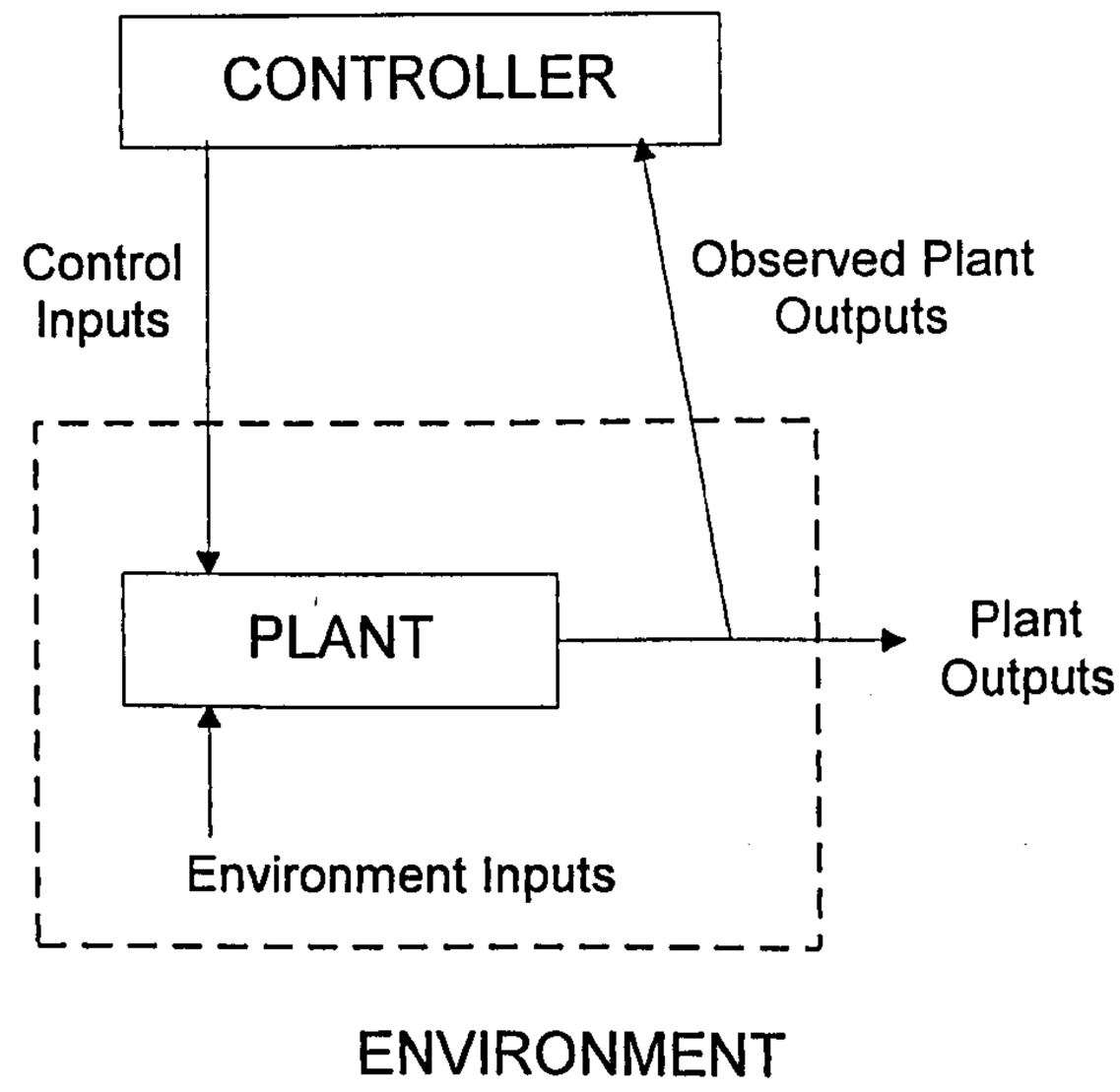


Figure 1.1 - Simplified Control Scheme

The design of the controller is usually achieved by applying a particular control paradigm to suit a particular task. An example of this is model-based control which identifies a model of the plant, and then uses this model to set the internal parameters of the controller. This approach has been widely adopted by conventional control systems in industry, and has found successful application in many tasks where the controller is required to maintain dynamical performance despite disturbances within the plant or the environment. However, these control systems are increasingly expected to cope with more complex tasks, such as when the dynamic processes in the plant are highly non-linear, or when observing and measuring the plant outputs involves uncertainty. It has been argued that as tasks become even more challenging and complex, the control systems used to control them will require a different approach [Brown & Harris, 1994]. Autonomous behaviour is one such challenge, where the control system must be able to perform well under significant changes and uncertainties in the plant or the environment for extended periods of time without external human intervention [Antsaklis et al., 1991]. In order to achieve this, the control system must be able to learn about the environment based only on observations of that environment, and use this information to adapt its behaviour to deal with any changes that may affect performance.

There are many tasks that require autonomy, e.g. control systems used in nuclear power stations, fire fighting vehicles, and in unmanned space exploration where a spacecraft may be beyond the direct control of an Earth-based operator because of communication delays [Gupta & Rao, 1994]. In these tasks, the autonomous controller is expected to deal with unexpected situations and new control tasks, while still being able to tolerate failures within certain limits. Such tasks may occur in situations where human involvement is either hazardous, tedious or impossible, and control can only be achieved by means of the information directly available to the control system. Autonomous control

of this type is the motivating factor in the field of *Autonomous Control* which has specifically attempted to address these difficult problems [Antsaklis, 1995].

Autonomous control is made possible because the control system is able to acquire knowledge about the state of its environment based upon its experiences in that environment. It then uses this knowledge to adapt its behaviour in response to situations that it might not have yet encountered, i.e. it possesses the ability to apply its existing knowledge in states where knowledge is yet to be acquired. A problem arises when the environment changes because the knowledge the controller possesses about particular states no longer correlates with the environment. The controller therefore needs to be able to identify when this situation occurs, and learn new knowledge about those states in the environment that have changed. This might involve having to learn about the whole environment all over again, although it is possible that some of the knowledge acquired by the controller may still be applicable to parts of the environment. This is the *relearning problem* because the controller needs to decide which parts of the environment need to be learned, and this decision could be based on what the controller perceives to be novel or unexpected aspects of either the environment or its own behaviour. Thus, the controller needs to possess mechanisms that allow it to :-

- Obtain knowledge about the environment and its actions based on observations of both the environment and its own behaviour
- Recognise when the knowledge it has acquired is appropriate to a particular situation in the environment, and use this knowledge to modify its behaviour
- Identify when the knowledge it has acquired is no longer appropriate to a particular situation in the environment, and thus obtain new knowledge about the environment.

The relearning problem is different to learning about the whole environment all over again because it allows for the learning of new knowledge simultaneously with the retention of still useful knowledge from past experiences, an ability not typically offered by existing control systems [Denham, 1994]. This issue must be addressed for the successful realisation of autonomous behaviour in control systems, and is one of the key aspects to this thesis. If the relearning problem is not adequately addressed, then costs could be incurred as a result of unnecessary learning even when only minor changes occur in the environment, and this is clearly not desirable in autonomous control systems.

Intelligent Control is one of the more recent approaches to dealing with the need for autonomous behaviour, as well as the problems of increased complexity and greater uncertainty in the plant and environment. Section 1.2 explains what is meant by 'Intelligent Control', and provides a detailed look at what this approach hopes to achieve. Section 1.3 looks at the imperatives that have led to a greater emphasis on biological control systems, and how an understanding of the underlying mechanisms provided by these systems can be considered a source of inspiration for further developments in intelligent control. Section 1.4 looks at the specific objectives of this research, and the approach taken towards achieving these objectives. Section 1.5 provides an outline of the structure of the thesis, and the contribution made to knowledge by particular aspects of this research.

1.2 - What is Intelligent Control ?

The term 'intelligent control' has been around for very many years, and Brown & Harris [1994] point out that the idea of intelligent control was originally proposed to extend the range and flexibility of then existing automatic control systems. There has since been much research activity in the field of so-called "Intelligent Control", with various attempts at defining both the term and the field of research [Denham, 1994]. There are two

major issues that need to be addressed. The first issue considers what is meant by the 'intelligent' aspect of intelligent control. This is a problem in itself because there has been no real consensus as to what defines intelligent behaviour in either living beings or artificial systems, and therefore this thesis will make use of the definition of an intelligent system used by Werbos [1994]:

"A system capable of maximising some kind of measurement of utility or reinforcement or performance or goal-satisfaction (with or without prior knowledge of how that measure is defined as a function of other variables) over time, in an environment whose dynamics are not known in advance, so that the system must learn both the dynamics and a strategy of action in real time through experience."

The second issue considers the role of control in an intelligent system. In May 1993, a Task Force working at the invitation of the Technical Committee on Intelligent Control of the IEEE Control Systems Society set about addressing both of these issues. The Task Force's main aim was to define what is encompassed by the term 'intelligent control', and its key objectives were outlined as follows [Antsaklis et al., 1994]:-

- To characterise intelligent control systems, thereby clarifying the role of control in intelligent systems
- To be able to recognise these intelligent control systems, and distinguish them from conventional control systems
- To help identify problems where intelligent control methods appear to be the only appropriate techniques.

The Task Force found that the area of intelligent control is in fact interdisciplinary, it attempts to integrate or extend theories and methods from areas such as conventional control, operations research and artificial intelligence in order to meet the demands of complex control problems. Intelligent controllers are seen as control systems that are able to emulate some of the capabilities of intelligent biological systems, such as adaptation and learning, planning under large uncertainty, and coping with large amounts of information. Antsaklis et al. [1994] state that this has always been used as the prime justification for the word "intelligent" in intelligent control because it is these capabilities that are considered to be the important attributes of human intelligence. They cite the following as examples of intelligent control systems on an extensive list of real-time control system implementations compiled by the National Institute for Standards and Technology :-

- NASA space-station telerobot
- Intelligent highway vision-based road following vehicle
- Autonomous undersea vehicle.

Topics and applications in the field of intelligent control are gradually evolving and extending the areas of conventional control systems, greatly assisted by recent advances in computing technology. It may be argued that two unique features differentiate intelligent control systems from conventional control systems: the ability to make decisions, and the ability to learn [Shoureshi, 1991]. At the same time, it may also be argued that there is no clear distinction between intelligent and conventional controllers because intelligent controllers often consist of both intelligent and conventional components [Passino, 1993]. The important point is that the integration of intelligent and conventional control

approaches hopes to achieve the capabilities that have up until now only been possible in control systems that were operated by humans.

1.3 - Biological Inspiration

The focus in intelligent control is on designing controllers that can perform or emulate certain functions of intelligent biological systems in order to solve control problems. It is logical to assume that developments in the field of intelligent control are limited by our current understanding of the fundamental processes that occur in biological systems, and for this reason existing intelligent controllers can only weakly reproduce the complex functions of their biological counterparts. Gupta and Rao [1994] argue that the incredible flexibility and adaptability of biological neural control mechanisms may provide the inspiration for developing more capable intelligent control mechanisms. Their argument is based on the fact that biological methods of processing information are fundamentally different from those used in conventional control techniques, but solve similar problems. For example, a robot arm that must pick up an object performs the same task as a human picking up that object. The human (i.e. biological system) executes the task at a conscious level, but subsequent computations are performed subconsciously e.g. muscle coordination and detailed calculations of joint angles are carried out in subconscious computing centres of the central nervous system. To perform the same task, the robot arm must measure the position of its hand relative to the object, and then compute the direction vector to move the hand towards that object. Gupta & Rao [1994] argue that this requires a great deal of computation and *a priori* knowledge about the robot's environment and the robot arm itself, such as the position of arm joints etc. They also argue that if either the environment or the robot arm changes, then a robot arm using a traditional control methodology may fail to perform the desired task. A human can still perform the

task because the biological processes can easily adapt to changes in the environment or changes to muscles and joint angles that might occur at a subconscious level. Therefore, if robots are to perform the same tasks as humans, they need to be able to emulate the capabilities found in biological systems. Biological systems thus provide a clue as well as a challenge for the design of artificial intelligent systems that can emulate capabilities for dealing with the uncertainty in executing complex tasks in an unstructured environment.

It is argued that emulating the precise neurophysiological behaviour of biological control systems is not necessary as it is sufficient to simply incorporate some of the computational operations that facilitate biological learning and adaptation [Gupta & Rao, 1994]. The basic hypothesis is that if the fundamental principles of neural computation used by biological control systems are understood, then this may provide inspiration for developing an entirely new generation of control methodologies that provide capabilities not found in existing control techniques. Werbos [1991] also argues that inspiration from the brain is important because the brain may be considered a living example of a controller capable of controlling millions of variables simultaneously under conditions of extreme non-linearity, uncertainty and noise. The last few decades have seen considerable progress in our understanding of neurobiological systems, and the most recent discoveries in neuroscience have greatly contributed to our understanding of the structure, function and neurochemistry of biological control systems. The rationale behind the research conducted for this thesis is that these discoveries need to be taken into account if artificial control systems are to effectively emulate the intelligent capabilities found in biological control systems. There are other researchers who have adopted a similar rationale. The pioneering work of Sutton & Barto [1981] was based on earlier work into the computational modelling of learning processes involved in biological systems, and was supported by evidence from learning and behavioural studies. More recently, Grossberg & Merrill

[1992] proposed a computational architecture modelled on a brain region called the hippocampus, and showed that this architecture is able to reproduce some of the intelligent functions thought to be attributed to this brain region.

1.4 - Objectives of this Research

The ability to adapt is a fundamental characteristic of biological organisms since they attempt to maintain physiological equilibrium in the midst of changing environmental conditions [Ogata, 1990]. This may be considered an adaptive process because the organism must go through the stages of learning about its behaviour in the environment, identifying changes to dynamic processes occurring in both organism and the environment, and then modifying its behaviour accordingly. The learning process is also fundamental to intelligent behaviour, and the computer modelling of the process of learning has been the subject of research in the field of *Machine Learning* for very many years. Learning in a control system may be defined as the process by which the controller can alter its actions to perform a particular task more effectively due to increases in knowledge related to the task [Antsaklis, 1995]. Learning is therefore an important feature of any intelligent or autonomous control system, and the introduction of learning methods into control has been an attempt to widen the range of existing control system applications. The objective of the research conducted here is to look more closely at the learning processes that occur in biological systems, as these processes are likely to be involved in autonomous behaviour and other intelligent functions. Specifically, this research will address the relearning problem and look at how biological systems deal with this problem. The aim is to see if there are mechanisms for dealing with the relearning problem inherent to biological systems that may provide inspiration for similar mechanisms in intelligent control systems.

1.5 - An Outline of the Thesis

The remainder of the thesis is organised as follows. *Chapter Two* provides a detailed look at learning control systems, describing how they are able to learn about the dynamic processes occurring in their environment, and how they are able to adapt to changes in that environment. A particular type of learning known as reinforcement learning will be introduced, and the significance of this type of learning will be explained. One of the more successful approaches using reinforcement learning in control systems has been the *Adaptive Critic* approach, a computational architecture that has been used to solve a variety of control problems. The adaptive critic system assumes that the controller will learn to control a given system based upon a *fixed schedule of reinforcement.* Reinforcement schedules are described in the chapter, but essentially this means that the adaptive critic system has been designed to learn about the environment under the assumption that the reinforcement information that it receives will always be representative of the environment. This reinforcement information is not expected to change and thus the adaptive critic system continually learns about the environment. This is only one aspect of the overall reinforcement learning problem, and has been described by Barto [1995] as a 'subproblem'. The adaptive critic system must be able to detect changes to reinforcement schedules in order to adapt its behaviour accordingly, i.e. know when relearning is necessary. This is the relearning problem because the adaptive critic system has already acquired knowledge during the initial process of learning, and some of this knowledge may still be useful and not need to be learned again. The chapter discusses the need for a mechanism that can detect changes in reinforcement schedules, and how this mechanism may be used to address the relearning problem. The implications of the relearning problem have not been considered in this context, and therefore the work in this thesis can be seen

as a contribution to existing research on learning control systems and reinforcement learning.

Modern reinforcement learning is based on the fundamental concepts derived from theories of animal conditioning, and has also been influenced by concepts in artificial intelligence and control theory. This has been the main inspiration behind the development of the adaptive critic system described in Chapter Two. Despite the similarity between the architecture used by this system and the structure and function of certain regions of the brain, relatively little effort has been made to relate this architecture to the nervous system [Barto, 1995]. The nervous system of all animals has a number of basic functions in common, including the coordination of movement and the analysis of sensory information [Llinas, 1990]. The nervous system may thus be considered a biological control system in that sensory information comes in from the environment (vision, sound, smell, etc.), and is then processed by specialised areas of the brain. The processed output is used to determine the future behaviour of the animal, and therefore the processing of sensory information must include an evaluation of the animal's previous behaviour. This constitutes a reward system because desirable behaviours are rewarded, and undesirable behaviours are punished. *Chapter Three* investigates the biological basis of the reward system that enables reinforcement learning in the brain, and looks in detail at the anatomy and physiology of the amygdala (the part of the brain thought to assign emotional significance to sensory stimuli) and the basal ganglia (the part of the brain thought to act as the interface between sensory and motor information processing). The objective is to develop a conceptual model to support the hypothesis that the amygdala influences learning in the basal ganglia, and that this influence is particularly important when relearning is necessary. The model is based on the involvement of various neurochemical substances found in these brain regions, and the functional implications of these substances are discussed. This work is an

attempt to bring together the findings from a number of recent studies into a coherent biological model of the reward system that facilitates reinforcement learning (and relearning) in the brain, and thus makes a contribution to the understanding of the biological mechanisms involved.

Having developed a conceptual model to explain the involvement of the amygdala in reinforcement learning and relearning, the aim of the thesis is to then show how this model can be used to provide established control approaches using reinforcement learning (e.g. the adaptive critic system) with mechanisms that can address the relearning problem. The conceptual model accounts for the interaction of the amygdala with structures in the basal ganglia. Houk et al. [1995b] have developed a model that attempts to relate the anatomy and physiology of structures in the basal ganglia to the theory of adaptive critics. *Chapter Four* describes the Houk et al. model, and shows how the basal ganglia can be considered functionally equivalent to the adaptive critic system. The chapter then describes how the adaptive critic system may be relocated in another part of the basal ganglia in accordance with the conceptual model presented in the previous chapter, yet still retain the functional characteristics of the adaptive critic system. This therefore allows the interaction of the amygdala to be considered, and how this can be used to modify the adaptive critic system to provide new capabilities. In the conceptual model, the amygdala modulates the effect of various neurochemical substances, and it is suggested that the role of these substances is equivalent to the role of the learning parameters in the adaptive critic system. The chapter attempts to show how these parameters can be related to neurochemical substances, and suggests how the adaptive critic system can be modified to include the functions of the amygdala in relearning. The computational modelling of the amygdala as a modulator has not previously been considered, particularly in the context of relearning. The modified adaptive critic system with modulation of parameters governed by an amygdala-

inspired mechanism therefore represents an original contribution to knowledge, and may lead to mechanisms that effectively address the relearning problem.

The pole balancing problem is a classic control task examined exhaustively in control theory texts, and is a good example of an inherently unstable system representative of a wider range of tasks [Wieland, 1991]. *Chapter Five* presents the pole balancing problem as an experimental framework for evaluating the success of a reinforcement learning control system, and describes how this problem needs to be extended to investigate the relearning problem. An experimental benchmark definition for simulation of the pole balancing problem is presented [Geva & Sitte, 1993], and the inclusion of the benchmark in this work is an attempt to standardise the problem for future comparison. *Chapter Five* uses the modified adaptive critic system described in Chapter Four to solve the benchmark pole balancing problem extended to include the relearning problem. The performance of the modified system is shown by providing a set of preliminary results based on simulation experiments. These results have been analysed, and compared to the original adaptive critic system working on the same problem. This work represents a contribution to knowledge because it provides empirical results that can be compared to future work on the relearning problem, which may lead to further developments in intelligent control and the objectives of this research being achieved.

Chapter Six provides a summary of the thesis and an outline of the contribution that this research has made to knowledge. In addition, the chapter looks at the main conclusions that can be drawn from this work, and outlines the limitations of the research conducted. The possibilities for future work arising from this research are discussed.

Chapter Two

Reinforcement Learning and the Relearning Problem

2.1 - Introduction

The ability to adapt is important to intelligent control systems, as discussed in Chapter One. The term 'adaptive' has a variety of specific meanings, but usually implies that a system is capable of modifying its own behaviour in response to disturbances in its environment, regardless of whether these are internal or external to the system [Ogata, 1990]. This chapter discusses learning control, a field in control that has developed because of the need for control systems that are able to adapt. The chapter then looks in detail at reinforcement learning, a sub-class of learning control that has achieved much success when dealing with the problems of intelligent control. Reinforcement learning is the basis for the *adaptive critic system*, a computational architecture that has received much attention in intelligent control because of its "brain-like" characteristics [Werbos, 1995]. The chapter discusses how the adaptive critic system successfully utilises the reinforcement learning framework to solve learning control problems, but is found lacking when it is required to address the relearning problem. The chapter explains what the implications of the relearning problem are in the context of the adaptive critic system.

2.1.1 - Feedback Control

All control problems involve manipulating the input to a dynamical system so that the behaviour of this system meets a set of pre-specified requirements that constitute the

control objective [Sutton et al., 1992]. The task of the controller is to determine control inputs in accordance with the control objective, and sometimes the control objective is specified in terms of target outputs that the output from the plant should match or track as closely as possible. This is known as *Feedback Control*, and assumes that the target outputs are known and can be supplied to the controller. The feedback control scheme is illustrated in Figure 2.1.

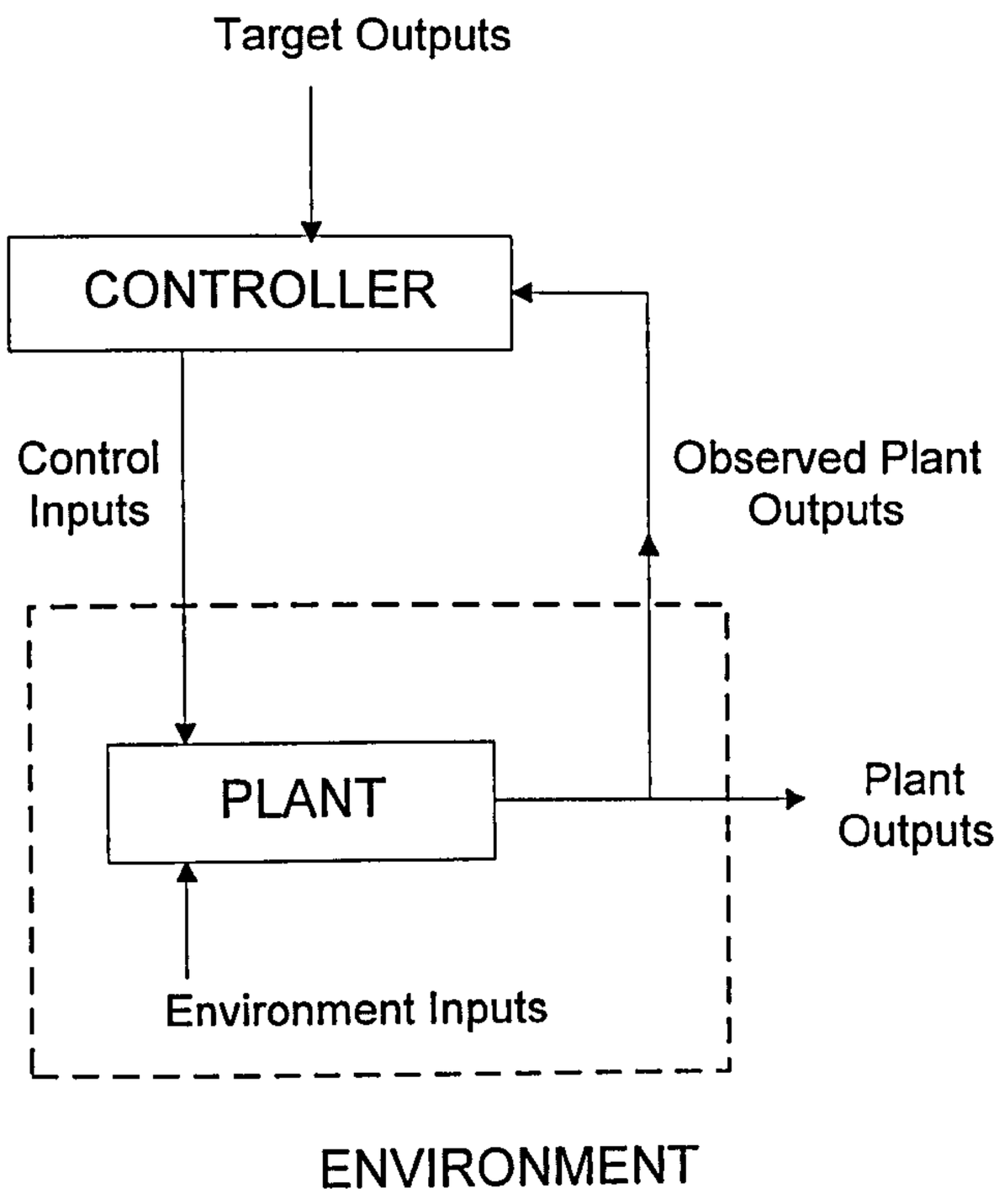


Figure 2.1 - Feedback Control Scheme

However, it is often the case that target outputs cannot be provided to the controller because disturbances in the plant or environment create uncertainty so that the provision of target outputs is impossible or extremely difficult. If the control system is viewed as a mapping from desired plant outputs to control inputs, then the information required to make this mapping needs to be available *a priori* so that the controller can be appropriately designed. If this information is in any way limited or inaccurate (such that disturbances

cannot be accounted for), then the controller will be unable to provide the required control inputs that produce the desired plant outputs. This is where *Adaptive Control* is used.

2.1.2 - Adaptive Control

An adaptive control system is one that can adjust itself to accommodate new situations, such as changes in the observed dynamical behaviour of the plant [Baker & Farrell, 1992]. The adaptive control system monitors the input/output behaviour of the plant to identify the parameters of an assumed dynamical model, and adjusts these parameters to determine control inputs to produce the desired plant outputs. The adaptive control scheme is illustrated in Figure 2.2.

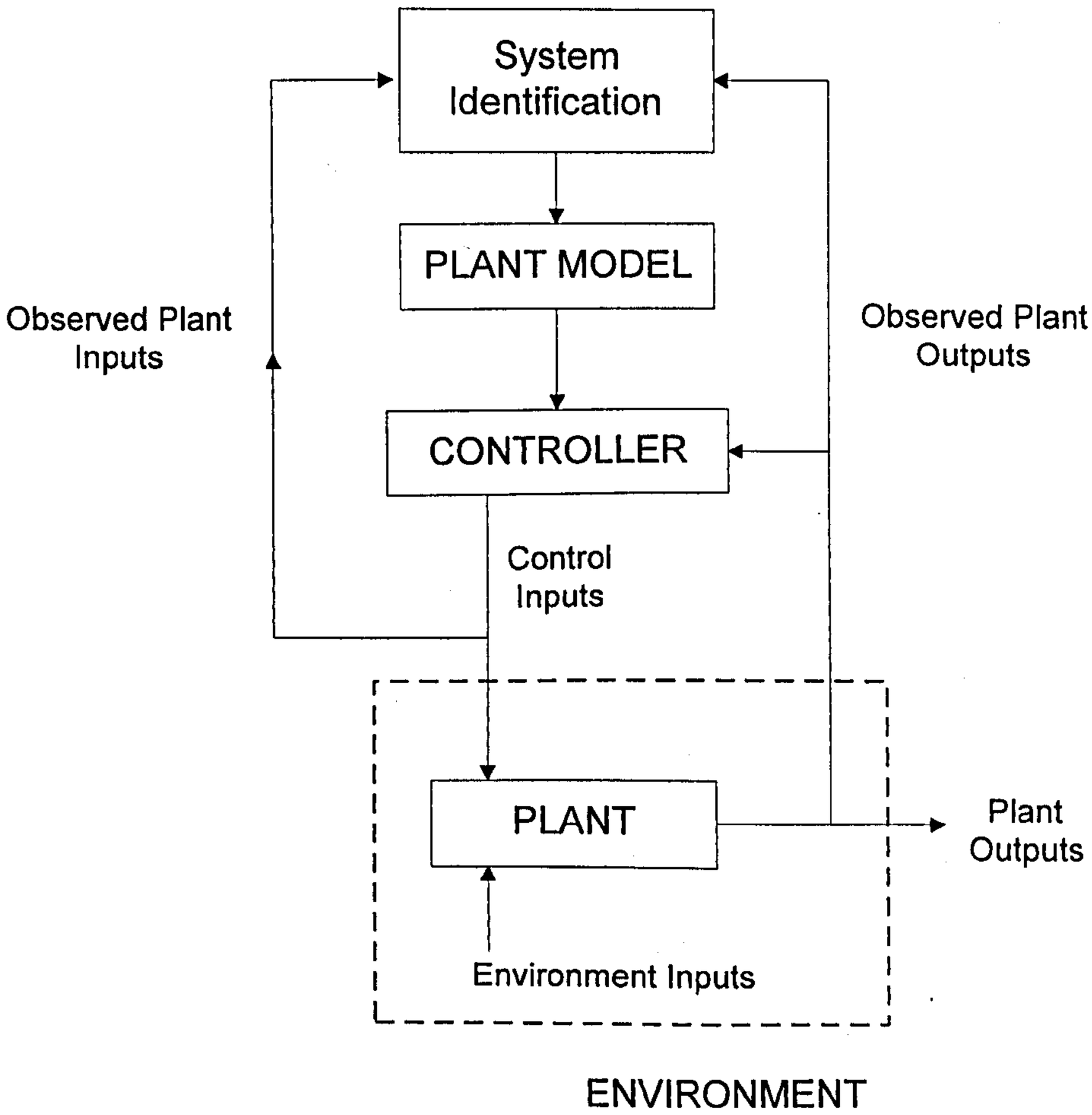


Figure 2.2 - Adaptive Control Scheme

An adaptive control system will attempt to adapt the parameters of the plant model whenever the behaviour of the plant changes by a significant degree, and this process is sometimes referred to as 'system identification'. Baker & Farrell [1992] argue that if the dynamical behaviour of the plant varies considerably over its operating envelope (e.g. due to nonlinearity), then the control system may be required to adapt continually. This is undesirable because degradation in performance can be associated with periods of adaptation. Such adaptation periods may also be required in the absence of nonlinearity or disturbances since the controller must re-adapt every time a different dynamical regime is encountered, and adaptation occurs even if the system is returning to an operating condition that it has encountered and handled before. This is clearly undesirable, and may incur unnecessary costs in terms of computational expense. If the control system could somehow use past experience to determine its control inputs, this would greatly reduce the computational burden on the control system. This is where *Learning Control* is used.

2.1.3 - Learning Control

A learning control system is one that has the ability to improve its performance in the future, based on experiential information it has gained in the past [Baker & Farrell, 1992]. This implies that the learning control system is in some way *autonomous* because it can improve its own performance, and that it has *memory* since it can exploit past experience to improve its future performance. Adaptive controllers lack "memory" because they must re-adapt to compensate for all apparent variations in the dynamic behaviour of the plant, but learning controllers are able to correlate past experiences with present situations thereby not treating every distinct situation as a novel one. The learning control system will adapt the parameters of the plant model in accordance with the memory of its experiences, and this constitutes the learning aspect of this type of control. The plant model

is then used by the controller to determine control actions in order to achieve the control objective more effectively due to increases in knowledge related to the task [Antsaklis, 1995]. This increased knowledge is contained in the plant model, but the control system also needs to have an index of performance in order to relate control actions to particular situations. This index is known as the *objective function*, and is used to evaluate the actions of the controller. The control system thus updates the plant model on the basis of the objective function. An accurate plant model means that when particular situations are known, appropriate control actions can be determined. The mapping from situations to actions is known as the *control policy*, and the controller must learn a policy that gives the maximum value for the objective function. The aim of learning control is therefore to use the objective function to find the best control policy, which is often referred to as the optimal control policy. The distinguishing feature of learning control is that it uses past experience to determine the best control policy, and this policy is manifested in the plant model. This approach has been successfully applied to numerous learning control problems, but has only been possible when the objective function is known [Gullapalli et al., 1994]. There are many control problems where the objective function cannot be explicitly expressed due to the complexity of the problem. For example, suppose the behaviour of the plant can only be improved in accordance with a performance measure that evaluates the overall behaviour of the plant, such as maximising a measure of the energy efficiency of the plant over time. This requires a learning control scheme known as 'reinforcement learning', where the objective function is not necessarily known and must be found. Section 2.2 describes the theory of reinforcement learning, and the way in which it addresses the problems of learning control. Section 2.3 outlines the adaptive critic approach, an architecture that uses the reinforcement learning control scheme and has already achieved much success. Section 2.4 discusses how the adaptive critic approach has

difficulties when it comes to dealing with the relearning problem, and suggests that inspiration derived from biological mechanisms of reinforcement learning and relearning may help resolve these difficulties. A better understanding of the biological mechanisms is consistent with the rationale behind this work, i.e. that this inspiration may lead to developing computational mechanisms better able to address problems such as the relearning problem. Section 2.5 provides a summary of this chapter.

2.2 - Reinforcement Learning

Reinforcement learning addresses the problem of improving performance as evaluated by any measure whose values can be supplied to the control system [Barto, 1989]. This may also be described as the problem faced by a control system that must learn appropriate control behaviour based only on trial-and-error interactions with the plant. The desired control signals are those that lead to optimal plant performance, but the learning system is not told what these are because this information is not available. This means that instead of trying to determine control inputs from target plant outputs, the control system tries to determine target control inputs (or desired changes to control inputs) that lead to increases in the measure of plant performance, which is not necessarily defined in terms of target plant outputs [Barto, 1989]. Reinforcement learning is therefore a sub-set of learning control because the problem is to find the optimal control signals, not simply to remember and generalise from them.

Reinforcement learning is not so much a learning method as a framework within which a wide range of learning control problems can be formulated. Sutton [1992] states that reinforcement learning is based on the notion that if an action results in a satisfactory or improved situation, then the tendency to reproduce that action is strengthened. The converse is also true, such that if an action results in a poorer situation, the tendency to

reproduce that action is weakened. Learning takes place by providing the control system with a reinforcement signal that either 'rewards' desirable situations, or 'punishes' undesirable situations. The reinforcement signal is a representation of the objective function because the control system is required to maximise the total reward (or minimise the total punishment) that it receives. The control system therefore relates its control actions (control inputs) to situations in accordance with the evaluation of the situation given by the reinforcement signal. The nature of the reinforcement signal is discussed in Chapter Three.

2.2.1 - The Value Function

Consider the problem faced by a mobile robot navigating through a maze. Each possible situation (state) in the maze that the robot finds itself in discrete time is given by x_t, where $x \in X$. The robot can perform an action $a \in A$ to move it from the state x_t to the next state x_{t+1}. The robot is required to reach a terminal state in the maze that has been predefined as the goal state, x_{goal}. The robot is given a reinforcement signal $r(x_t)$ in each state, except for the starting state x_0. This signal punishes the actions of the robot by making $r(x_t) = -1$ until the goal state is reached, at which point the reinforcement signal $r(x_t)$ is set to zero. The control system is required to sum the total reinforcement over time received by the robot until it reaches the goal state x_{goal}. This summation is calculated by the *value function*, $V(x_t)$ given by Equation (1):

$$V(x_t) = \sum_{k=0}^{\infty} r(x_{t+k}) \qquad (1)$$

$$= r(x_t) + r(x_{t+1}) + \ldots + r(x_{goal})$$

In Equation (1), k represents time steps in the future, and x_{goal} is the first goal state the agent encounters after time t assuming the robot stops when it reaches the goal. The robot is encouraged to reach the goal as quickly as possible by maximising $V(x_t)$, which therefore gives the highest possible value. It should be noted that the reinforcement signal represents only the overall objective of the control system because all states except the goal state are punished. The controller does not specify *how* the objective should be achieved, i.e. the exact path to be taken. This means that the control system may focus only on the control objective which is to reach the goal state as quickly as possible. Otherwise, it is possible that the control system will find a way to achieve subgoals (reach other states that are not punished) without ever achieving the overall control objective.

Reinforcement learning finds a control policy for mapping from situations to actions based on the reinforcement signal. If the control actions are denoted by a (where $a \in A$), and the states are denoted by x (where $x \in X$), then the policy π for mapping between x and a is given by $\pi : X \rightarrow A$. The controller is thus an implementation of the control policy because it maps states to actions in accordance with the policy. If the reinforcement signal is not provided externally, then the control system needs some kind of internal mechanism that will indicate which states are (or are not) desirable. This is referred to as the *critic*, which generates internal reinforcement signals based on observations of the plant. The critic observes the outputs from the plant (in the absence of a reinforcement signal), and then provides the controller with an internal reinforcement signal that reflects the success of its control actions. The control system designer determines the way in which the internal reinforcement signal is produced, such that it evaluates the actions of the controller and how well the controller achieves the desired behaviour of the plant. The control system uses the internal reinforcement signal to determine an appropriate control

policy which will maximise the reinforcement that it receives. Figure 2.3 is an illustration of a reinforcement learning control scheme using a critic.

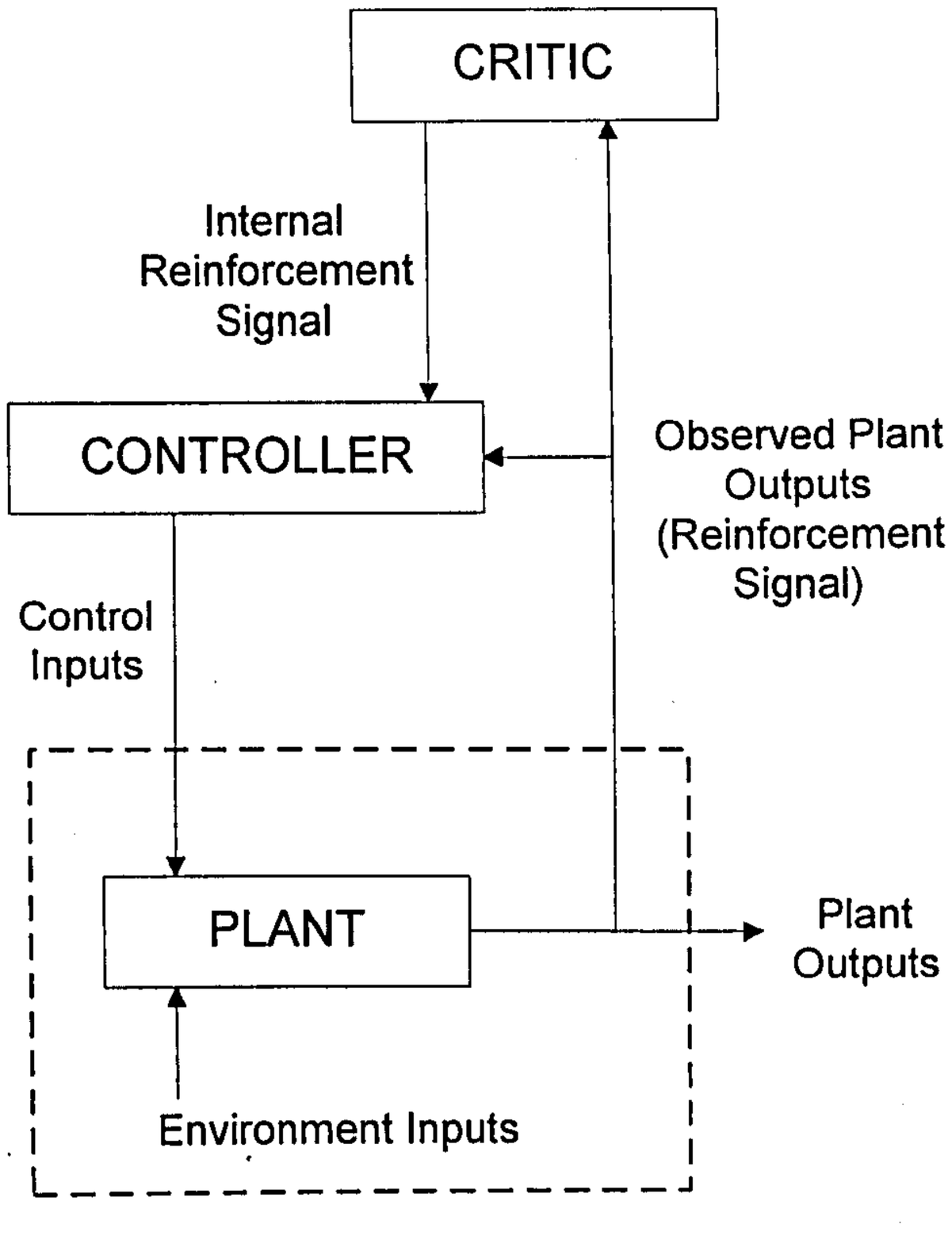


Figure 2.3 - Reinforcement Learning Control Scheme

Therefore, it can be seen that a learning control problem may be framed as a reinforcement learning problem in terms of :-

- The choices made by the controller (control actions)
- The situations in the plant or environment on which the choices are made (states)
- The criteria which defines the control objective (reinforcement signal).

The design of a reinforcement learning control system essentially involves two problems [Barto, 1990]. The first problem is to construct the critic so that it is capable of evaluating the performance of the plant appropriate to the control objective, and allows the learning of an optimal control policy. In other words, a value function must be learnt that can rapidly map states to values. The second problem is to determine how the control signals should be modified to improve the performance of the plant based on the internal reinforcement signal provided by the critic. In other words, a control policy must be learnt that can rapidly map states to actions. These problems are addressed separately by the two main components in this type of reinforcement learning system, called the 'actor-critic' architecture for reinforcement learning control. In this architecture, the actor is responsible for learning the control policy for control signals to the plant, and the critic is responsible for the value function that determines the internal reinforcement signal. This architecture is the basis of the adaptive critic approach discussed in Section 2.3. There are a number of aspects to the design of a reinforcement learning system which are outlined in the following sections. Much of the material in these sections is taken from the work of Kaelbling et al. [1996], Sutton & Barto [1995], and Barto [1995].

2.2.2 - The Environment

Consider the problem of a reinforcement learning control system that chooses a particular control action a by sending a control signal to the plant. The plant uses this to change the state x depending on the dynamic processes and disturbances to the plant and the environment. The critic then provides an internal reinforcement signal based on the consequences of taking the action a in state x. This problem can be modelled as a *Markov Decision Problem* (MDP) because the internal reinforcement signal provided by the critic depends only on the consequences of taking action a in the current state x, and not on any

previous actions or states. An MDP consists of a set of states X, a set of actions A, a reward function R: $X \times A \rightarrow \Re$, and a state transition function T: $X \times A \rightarrow \Pi(X)$ where $\Pi(X)$ is a probability distribution over the set X. The state transition function T maps states to probabilities, and this gives the probability of making a transition from the current state x_t to the next state x_{t+1} using the action a in discrete time. The reward function R specifies the expected reward as a function of the current state x_t and the action a. The model is *Markov* if the state transitions are independent of any previous state or action.

2.2.3 - The Discount Factor

The objective of the learning control system is to find a policy π that gives the maximum $r(x_t)$ value for each x_t, which is referred to as the *optimal policy*. An optimal policy is any policy that is "greedy" with respect to selecting actions that maximise the overall reinforcement that is obtained. There may be several optimal policies, but all of these policies share the same optimal value function. If the goal state is denoted by x_{goal}, and the robot is given a reinforcement signal $r(x_t)$ in each state x_t, then the optimal value function is calculated by summing the total reinforcement received until the goal state is achieved when following an optimal policy. The *optimal value function* is denoted by $V^*(x_t)$, given by Equation (2):

$$V^*(x_t) = \max_{\pi} E\left(\sum_{k=0}^{\infty} r(x_{t+k})\right) \tag{2}$$

$$= \max_{\pi} E\left(r(x_t) + r(x_{t+1}) + \ldots + r(x_{goal}) \right)$$

In Equation (2), k represents time steps in the future, and x_{goal} is the goal state, and E() is used to indicate that the equation will yield the maximum expected value when the

optimal policy is used. The optimal value function accumulates the total reinforcement over time when the controller uses policy π starting at time t. This is based on the reinforcement signal, and equates to the maximum possible reward (or least possible punishment) that the system will receive. In some cases however, the reinforcement signal is delayed and may only be given at the end of a long sequence of control actions. This is known as the *temporal credit assignment problem*, and requires that the learning control system must determine which actions from a series of control actions deserve the credit (or blame) for improvements (or decrements) to the performance of the system. This problem can be solved by using a *discount factor*, 'γ', which determines the present value of future reinforcements [Sutton & Barto, 1995]. A reinforcement that is received k steps in the future will only be worth γ^k of what it would be worth if it was received now. The discount factor is included in the calculation of the value function $V(x_t)$, which is now given by:

$$V(x_t) = \sum_{k=0}^{\infty} \gamma^k r(x_{t+k}) \tag{3}$$

$$= r(x_t) + \gamma r(x_{t+1}) + \gamma^2 r(x_{t+2}) + \ldots$$

When $\gamma = 0$, the reinforcement from any state is just the immediate reinforcement from the transition to that state, and the optimal value function will give the maximum immediate reinforcement. As γ increases towards one, future reinforcements become more significant in determining optimal actions such that $\gamma = 1$ is the undiscounted case and all future reinforcements are taken into consideration. The value of the discount factor is therefore important, and this issue will be revisited in Chapter Four.

2.2.4 - Approximating the Value Function

There are a number of methods that can be used to find optimal control policies using discounted rewards [Barto et al., 1995]. Dynamic Programming (DP) methods are a family of algorithms that allow the discounted reward case of the optimal value function to be written in a special form known as the *Bellman Optimality Equation*, which is given by:

$$V^*(x) = \max_a E\left(r(x_t) + \gamma \; V^*(x_{t+1}) | x_t = x, a_t = a\right) \qquad (4)$$

Equation (4) is independent of any specific policy, and can be used to find an optimal policy by defining a system of equations (one for each state) that can be solved uniquely for $V^*(x)$ providing the dynamics of the plant are known. For each state x, there will be one or more actions that will obtain the maximum value according to the Bellman equation. These are all equally good control actions, and any policy that selects greedily from among these actions can be considered an optimal policy. DP methods can therefore solve the delayed reinforcement learning problem, but are equivalent to an exhaustive search that must consider all the possibilities, compute the probability of their occurrence, and evaluate their utility in terms of expected reinforcement [Sutton & Barto, 1995]. This is only made possible if the dynamics of the system are completely known, and assumes that there are enough computational resources to complete the computation. Sutton & Barto [1995] point out that this is never completely true in practice, and that it is better to *approximate* solutions using DP methods. This is a topic of ongoing research, and many different methods have been proposed that enable this to be done. These include Heuristic Dynamic Programming [Werbos, 1992] which defines a utility function that estimates the optimal "reward-to-go", and uses this to select a control policy which is automatically the

optimal policy. Another method is Incremental Dynamic Programming which uses an algorithm known as Q-Learning [Watkins, 1989]. The Q-Learning algorithm is very popular because it unifies the functions of the 'actor-critic' architecture, and is therefore easier to implement [Kaelbling et al., 1996]. The Q-function is given by $Q^*(x,a)$, and gives the expected discounted reinforcement received by taking an action a in state x. If $V^*(x)$ represents the optimal value function over all states x, then $V^*(x) = \max_a E\ (Q^*(x,a))$ if the Q-function continues to select the best actions. The Q-function makes the action explicit, and so the optimal policy is found by acting greedily to select the actions that have the maximum Q-value for each state. Q-Learning will be briefly revisited when exploitation and exploration issues are discussed later in this chapter.

DP methods exploit the fundamental principle of optimality, but are often criticised because of the extreme computational requirements that arise from conventional implementation using quantised states and actions [Millington & Baker, 1990]. This means that there have been few practical successes using reinforcement learning in large-scale, complex real-world problems. These problems still need to be resolved. Tesauro [1995] states that one of the more promising developments that may lead to overcoming such problems is the use of *Temporal Difference* methods.

2.2.5 - Temporal Difference Methods

The DP methods described in the previous section are most applicable to off-line learning because of their vast computational requirements, and the assumption that the problem can be modelled as an MDP. Temporal Difference (TD) methods are a family of algorithms that are less restricted, and are able to use on-line reinforcement learning with an incompletely known system to control the behaviour of that system [Sutton, 1988]. This is achieved by 'predicting' the future behaviour of the system. A *prediction* is an estimate

of future behaviour based on what happened in similar situations in the past. This can also be considered as the discounted estimation of the value function $V(x_t)$ which was given in Equation (3). The prediction is likely to be a good estimate of the value function because it incorporates a sample of the immediate reinforcement $r(x_t)$ [Kaelbling, 1996]. The basic idea behind TD is that learning is based on the difference between temporally successive predictions, and so the goal of learning is to make the current prediction more closely match the prediction at the next time step [Tesauro, 1995]. For example, suppose we would like to compare the prediction at two successive time steps, and these predictions are denoted by P_t and P_{t-1}. The prediction P_t is the estimate of $V(x_t)$ for a given state x_t, and is given by:

$$P_t \approx V(x_t) = r(x_{t+1}) + \gamma\ r(x_{t+2}) + \gamma^2 r(x_{t+3}) + \ldots + e_t \qquad (5)$$

In Equation (5), $\approx$ means "approximately equal", and e_t is the error in the prediction at time t. The prediction P_{t-1} is the estimate of $V(x_{t-1})$ for the previous time step, as given by:

$$P_{t-1} \approx V(x_{t-1}) = r(x_t) + \gamma\ r(x_{t+1}) + \gamma^2 r(x_{t+2}) + \ldots + e_{t-1} \qquad (6)$$

The term e_{t-1} is the error in the prediction at time *t-1*. The difference between the two predictions at adjacent time steps is called the 'temporal difference error' (or *TD error*) [Sutton, 1988]. This error is often referred to as *effective reinforcement*, and is denoted by $\hat{r}_t$. TD methods attempt to predict the next reinforcement signal based on the immediate

reinforcement (if any), plus the difference between the previous prediction and a discounted version of the current prediction. This can be summarised as:

$$\hat{r}_t = r(x_t) + \gamma \; P_t - P_{t-1} \tag{7}$$

If $\hat{r}_t$ is positive, then the effective reinforcement represents a move from a predicted bad state to a predicted better state. If $\hat{r}_t$ is negative, then the effective reinforcement represents a move from a predicted good state to a predicted worse state. Learning therefore attempts to maximise the effective reinforcement until $\hat{r}_t$ (the error) is zero, at which time the prediction is equal to the actual reinforcement that the system receives. This method is very simple computationally, and actually converges to the optimal value function if given sufficient time [Barto, 1995]. The predictions therefore replace the estimated value functions given in earlier equations, and can be used to learn optimal control policies through predicting and observing the consequence of successive actions. Learning stops when the TD error becomes zero, i.e. when all predictions become equal to the reinforcements actually observed, and the control policy is optimal.

The TD method described above still requires a large number of computations, but the algorithm can be modified so that it is suitable for learning concurrently with real-time system operation [e.g. Barto et al., 1995]. The method described is actually an instance of a more general class of algorithms developed by Sutton [1988] known as *TD(λ)*. The general TD(λ) can be calculated using Equation (8):

$$\hat{r}_t = \left(r(x_t) + \gamma \; P_t - P_{t-1}\right)\bar{a}(x_t) \tag{8}$$

In Equation (8), $\bar{a}(x_t)$ represents an *eligibility trace* which is kept for every state. This means that the effective reinforcement is calculated according to the eligibility of the current state the system is in, this eligibility being the degree to which the state has been visited in the recent past. When a reinforcement is received, the trace is used to update all states that have been recently visited according to their eligibility. There are various methods that can be used for keeping eligibility traces in every state x, one version of which is given by:

$$\bar{a}_{t+1}(x) = \lambda \bar{a}_t(x) + (1 - \lambda) \tag{9}$$

Equation (9) holds an eligibility trace for every state x, and decays with time in accordance with the value of the coefficient 'λ'. When $\lambda = 0$, the system will only look ahead one step when calculating the effective reinforcement, which is identical to the method used in Equation (7). When $\lambda = 1$, this is equivalent to looking ahead an infinite number of steps, and this converges to a solution much faster than by simply using $\lambda = 0$. However, this is likely to be computationally more expensive. Therefore, it is clear that the value of 'λ' needs to be carefully chosen if an improvement in computational efficiency is to be achieved, as compared to simply using $\lambda = 1$ or $\lambda = 0$. The issue of λ will be revisited in Chapter Four.

2.2.6 - Exploration and Exploitation

The previous sections have described how the optimal value function is the solution to a set of equations defined by the Bellman Optimality Equation, given in Equation (4). Learning can be seen as the process of improving the approximation of the optimal value

function by incrementally finding a solution to this set of equations. The Bellman Optimality Equation is defined over all states, and therefore dynamic programming requires that the optimal value function is satisfied for all states in the problem space. Similarly in Q-Learning, the Q-values will slowly converge to optimal as long as state-action pairs are tried often enough [Kaelbling et al., 1996]. This requirement introduces the need for experience, which involves exploration of the state space in order to gain knowledge about the environment. Consider the example of the robot in a maze. The robot needs to *explore* and gain knowledge about its environment as this knowledge will help to minimise the time the robot takes to learn a path through the maze. The knowledge gained also needs to be *exploited* so that the robot can minimise the costs associated with learning, such as negative rewards when the robot collides with obstacles. The robot does not know which actions will result in collisions until all of the state space has been explored. Although it is possible that the robot can learn a policy that is "sufficiently" good without having to explore the whole state space, there is still an important trade-off between exploration and exploitation. The robot needs to be able to efficiently explore the environment to maximise the effects of learning, but exploit the knowledge it gains to minimise the costs of exploration. Thrun [1992] argues that exploitation is part of efficient exploration because exploiting the knowledge gained constrains the system to the more relevant parts of the environment thus reducing learning time. However, he also argues that exploration is part of efficient exploitation because the costs of learning cannot be minimised over time without efficiently exploring the environment. There are various exploration techniques that can be used in reinforcement learning, and these fall into two major categories :-

- Undirected techniques - these are based on randomness (e.g. random exploration based on probability distributions) without considering previous experience
- Directed techniques - these use exploration-specific knowledge from previous experience to guide exploration and maximise the knowledge gained.

Thrun [1992] argues that directed techniques are superior because they make use of previous experience in the environment to guide exploration, and this is effective exploitation. Examples of directed techniques include :-

- Counter-based - counts the occurrence of states and then drives the system to less explored states, i.e. "go to the least visited neighbouring state"
- Recency-based - drives the system to least recently visited (or unvisited) states, and therefore maintains a record of the time elapsed since visiting a state
- Dynamic switching - uses a trade-off parameter with function thresholds to guide exploratory behaviour such that the "desire to explore" is balanced against the "desire to exploit"
- Selective attention - uses dynamic switching, but accounts for earlier decisions using a gain parameter that biases the system towards either exploration or exploitation.

More detail about these techniques can be found in Thrun [1992]. The main interest in these techniques lies in their close relationship to the relearning problem. The relearning problem considers the problem of how knowledge that already exists can be used to guide effective exploration of the environment as a result of changes to the environment, and how exploiting that knowledge can lead to minimising the costs of exploration. The selective attention technique uses a gain parameter to dynamically switch between exploration and

exploitation on the basis of previous experience. What constitutes this "previous experience" is an important issue as the work reported by Thrun [1992] describes selective attention as a means of maintaining equilibrium between exploration and exploitation. The relearning problem is not so much concerned with maintaining equilibrium, but instead considers how the appropriate balance between exploitation and exploration can make use of past experience in the environment. Note that the idea of selective attention from a biological perspective will be discussed in Chapter Three, and the use of gain parameters will be discussed in Chapter Four.

There are a number of methods that attempt to use experience effectively to explore the environment, and these often require a model of the environment. Real-Time Dynamic Programming [Barto et al., 1995] uses Q-Learning and a learned model to concentrate computational effort on the state spaces most likely to be occupied. This method is specific to problems where an agent has a particular goal state to achieve, and reinforcements are zero elsewhere. Prioritised Sweeping [Moore & Atkeson, 1993] and Dyna-Q [Sutton, 1990] also use a learned model to speed up the process of temporal credit assignment. Dyna-Q is of particular interest because this system has been tried on problems with robots in changing worlds. Dyna-Q has been used on the blocking problem (an obstacle is added that blocks the optimal path, and the robot is required to find a new optimal path), and the shortcut problem (an obstacle is removed to create a more optimal path, and the robot is required to learn this optimal path). Sutton [1990] reports that by keeping track of experience using an internal model, and then providing an exploration bonus to exploit this experience, a dramatic improvement can be made to the speed of learning.. The work in this thesis is related because the relearning problem is a generalised equivalent to the blocking and shortcut problems, and the research conducted has the same objectives i.e. how can existing knowledge be used to minimise the costs of learning.

2.3 - The Adaptive Critic Approach

The reinforcement learning method previously described was developed because of the need for adaptive capabilities in intelligent control systems. This need has resulted in a great deal of effort directed towards developing more brain-like control systems because these adaptive capabilities are intrinsic properties of the brain. Werbos [1995] argues that if intelligent control system designs are ever to be considered truly "brain-like", they must not only demonstrate applied engineering functionality, but also possess all three of the following adaptive components :-

(1) an *Action* or *Motor* system capable of generating control signals for the plant or environment

(2) a *Critic* or *Evaluation* system used to assess the long-term costs and short-term benefits of alternative actions

(3) an *Expectations* system which identifies and serves as a model of the external environment or plant to be controlled.

The first two components (the Action and Evaluation systems) are the same as the actor and critic found in the actor-critic architecture used for reinforcement learning described earlier. Werbos [1995] argues that a third component (the Expectations system) is essential in order to explain the results of animal conditioning experiments that have been the basis for modern reinforcement learning. The first two components are already part of the *Adaptive Critic* system, which Werbos [1995] describes as the only type of design that anyone has ever formulated (in engineering, biology or anywhere else) with any hope of explaining the generic kinds of capabilities seen in the brain. The theory of adaptive critics is a highly complex field of study with several levels starting from very

simple designs, and extending all the way up to the brain itself. It is appropriate to briefly describe these levels, because each level corresponds to the chronological and progressive development of the adaptive critic design :-

- *Level Zero* - the first adaptive critic design originally formulated by Widrow et al. [1973], and now obsolete in terms of its practical application to control systems.

- *Level One* - the Barto et al. [1983] design which uses an internal reinforcement signal to train the actor, and TD methods to adapt the critic. It is still widely used even though it learns very slowly when there are a large number of continuous variables. It has proven to be very robust in problems where there are fewer variables that are binary rather than continuous e.g. the pole balancing problem. The application of the level one adaptive critic design to the pole balancing problem is described in detail in Chapter Five.

- *Level Two* - Werbos [1995] describes these as "advanced" adaptive critics developed between 1990 and 1993. These designs use an Action-Dependent Adaptive Critic (ADAC) that sends derivative signals back to the actor, and use a backpropagation mechanism to adapt its parameters. The rich feedback received by the actor makes it possible to more effectively control a large number of variables.

- *Level Three* - this was the first real attempt to move into the realm of "brain-like" control. These use Heuristic Dynamic Programming [e.g. Santiago & Werbos, 1994] to adapt the critic, and the backpropagation of derivative signals through a *model* (the Expectations system) to adapt the actor.

- *Levels Four and Five* - these use more powerful techniques to adapt the actor and critic, particularly Dual Heuristic Programming (DHP) and its variations [Prokhorov & Wunsch II, 1996]. These techniques were designed to minimise the error in the derivative signals sent back to the actor. Santiago & Werbos [1995] reported early simulations of DHP, an optimisation technique that was explicitly designed to scale to very large numbers of controls. They argued that existing optimisation designs are limited to batch, off-line, and small engineering problems where only a few controls can be used. The human brain routinely handles complex problems requiring many controls in real-time, and so this capability should be provided by an intelligent control system. The DHP controller has four components: an Action network, Critic network, Model network, and Utility function. A detailed discussion of this design is beyond the scope of this thesis (for more details see [Prokhorov et al., 1995]). The important functional capability in the design is the backpropagation of values from the outputs to the inputs, and DHP provides a dual mechanism to achieve this.

Werbos [1995] proposes that the human brain might be a level five adaptive critic system, made up of three highly interconnected components. He suggests that new evidence supports the idea that the learning part of the brain is in fact made up of three entire adaptive critic control systems, which he describes as *subbrains*. The "upper brain" is the decision-making system, the "middle brain" is a task executor and implementer of intentions, and the "lower brain" is a high-speed motion co-ordinator for motor actions. Werbos cites a number of studies that point to parts of the basal ganglia as highly involved in providing reinforcement signals that are learned in a way that is remarkably similar to the TD methods described earlier. This issue is discussed in Chapter Four. Werbos argues

that more research needs to be done to better understand and exploit the recent findings from neuroscience, and the work presented in this thesis is a part of that effort.

The level one adaptive critic design will be the focus of effort in the remainder of this thesis. This design is already well understood, but little effort has ever been made to relate this to the corresponding neurophysiological mechanisms [Barto, 1995]. By considering the original level one design, it is hoped that some of the mechanisms suggested in this thesis will also be applicable to higher level designs. These higher level designs are advanced only insofar as they provide functions thought to be "brain-like". For example, level three introduces an expectations component into the design that allows better approximations for dynamic programming, but has the same basic components as the level one design. The work in this thesis does not devalue the significance of the higher level designs, but takes the position that recourse to the biological mechanisms needs to be understood at a lower level before it can be applied at the higher levels. Indeed, it may be argued that backpropagation and similar mechanisms introduced at higher levels are not really biologically plausible given our current knowledge about the brain [Denham, 1994]. This position is consistent with the rationale of this research because as more is understood about the structure and function of the brain, perhaps this knowledge will support or lead to a better understanding of the higher level designs.

2.4 - Adaptive Critics and the Relearning Problem

The adaptive critic approach relies on the fact that the learning control system operates under a fixed schedule of reinforcement reflected by the reinforcement signal that rewards or punishes the actions produced by the controller. Catania [1970] describes a reinforcement schedule as:

"...the conditions under which a response can produce a reinforcer. These conditions may include the time elapsed since some prior event, the number or temporal patterning of prior responses, or any variety of modifications and combinations of such specifications. When a particular schedule operates, these conditions determine the way in which reinforcement comes into contact with behaviour and generates a characteristic performance."

Learning in the adaptive critic system is dependent on the reinforcement schedule, and the controller eventually learns to predict the reinforcement schedule and develop a set of control actions (control policy) that optimises the reinforcement that it receives. It is assumed that the learning system will employ the same control policy throughout the prediction process [Barto, 1995]. This does not mean that the actor always produces the same action, but that it always responds in the same way if a particular situation recurs. This is because the prediction of reinforcement is directly related to the reinforcement schedule, and the control policy is derived from the predictions. The critic generates an internal (effective) reinforcement signal by learning to predict the expected schedule of reinforcement, which may be delayed or even absent. The internal reinforcement is used by the actor to modify control actions. This therefore demands that the schedule of reinforcement does not change, and provides an accurate representation of the control objective throughout the learning of the problem space. This may not always be the case, and thus leads to a situation where the relearning problem is a factor. The consequence of changing the reinforcement schedule while the control system is learning about the problem space can amount to changing the overall control objective, thus affecting the control policy. If this occurs, the control system (both actor and critic) will have already acquired some knowledge about the problem space, and this knowledge may still be

appropriate to particular regions of that space. The time required to change from one policy to another might be an important consideration for the intelligent control system, as might the learning costs (negative rewards) associated with responding to that change. There needs to be some mechanism that can detect when a change to the reinforcement schedule has occurred, can decide which states need to be relearned, and then direct computational effort to these states. This is the same as exploiting existing knowledge to maximise exploration, hence minimising both learning time and exploration costs. It would seem natural to assume that this is a selective process based upon previous experience with the schedule of reinforcement, and subsequent chapters in this thesis will consider how this can be achieved using the inspiration of biological systems.

The relearning problem has not been addressed with specific reference to the adaptive critic system using inspiration derived from biological systems. The work presented in this thesis attempts to do exactly that. There have been some studies that have touched upon surrounding issues, but little work has been done specifically on the relearning problem. For example, Anderson [1989] highlighted the importance of a mechanism in the adaptive critic system that can concentrate on appropriate regions of the problem space as and when required. He used artificial neural networks to represent the actor and critic in a level one adaptive critic system, and applied this to the pole balancing problem. Neural networks were used because they are able to generalise across states, and can therefore acquire information about states that have not yet been experienced because of this generalisation. However, in many real-world situations, the control objective does not need to be generalised to all states, but to only a small subset of states. Take for example a control objective that may be expressed in terms of avoiding punishments rather than receiving rewards. When relearning is necessary, the control system needs to be aware of states that already possess information about punishments, information which may still

be relevant. There would be little need to learn about the entire state space all over again. Such a mechanism would help reduce the computational requirements imposed on the system, thus minimising the costs associated with learning, and would be invaluable when reinforcement schedules change. Such a mechanism would therefore be required to detect changes to the expected schedule of reinforcement, identify where existing knowledge is still appropriate, and thus constrain exploration efforts to only the relevant states. This work is a contribution towards the provision of such a mechanism.

2.5 - Summary

This chapter looked at learning control systems, and introduced reinforcement learning as a framework that can be used to solve the problems of learning control. The adaptive critic system was introduced as an approach that has adopted the reinforcement learning framework, and leads the way towards more brain-like intelligent control. The relearning problem was described as a problem that intelligent control systems (perhaps using the adaptive critic design) need to be able to address in order to maximise existing knowledge and minimise the costs associated with learning. This ability is currently not found in the adaptive critic system, and it is suggested that a mechanism that can detect changes to the reinforcement schedule and use this knowledge to guide learning behaviour may lead to the provision of such an ability. An investigation into how biological systems deal with reinforcement learning and relearning may provide inspiration for developing artificial mechanisms similar to those in biological systems, and may eventually lead to intelligent control systems that possess the necessary adaptive capabilities for dealing with the relearning problem. The next chapter looks more closely at biological reinforcement

learning processes, and a conceptual model of reinforcement learning and relearning is presented based on a number of neurophysiological studies.

Chapter Three

A Model of Biological Reinforcement Learning and Relearning

3.1 - Introduction

This chapter focuses on biological mechanisms that are the basis for reinforcement learning and relearning processes in the brain, and outlines the neural structures that are thought to be involved. The function of these structures is considered, and a conceptual model proposed that attempts to explain how biological relearning mechanisms are provided. The model is based on the interactions between neural structures that make up the limbic and motor systems of the brain, and the main neurochemical substances involved. The model attempts to explain how the brain is able to use predictions of reinforcement to learn appropriate behaviour (i.e. how associations between perceptual stimuli and motor responses are formed on the basis of predicting external reinforcement). These predictions would no longer be appropriate when reinforcement schedules are changed, and thus detecting such changes could be useful for relearning processes. The conceptual model is based on the hypothesis that an area of the brain called the amygdala is involved in detecting when predicted reinforcements no longer correlate with actual reinforcements. The amygdala is then able to use this knowledge to modulate several neurochemical systems involved in the processes of learning, memory and attention when relearning is required. The conceptual model is derived from observations and experimental findings drawn from a number of behavioural studies, and this model is an attempt to put these observations into a coherent conceptual framework.

3.2 - Reinforcement Learning from a Biological Perspective

The previous chapters explained why it is important for intelligent control systems to have the ability to detect changes in reinforcement, and know how to exploit this knowledge when relearning is necessary. It would now be appropriate to examine the role that reinforcement learning (and relearning) plays in the control of behaviour in biological systems. Many of the key ideas in modern reinforcement learning have been derived from theories of animal conditioning, such as classical and instrumental conditioning.

3.2.1 - Classical Conditioning

Classical conditioning is a well studied phenomenon, and the neural mechanisms involved in classical conditioning have been extensively investigated [Carlson, 1986]. This form of conditioning can be viewed as a mechanism that allows an organism to make predictions about the reinforcing properties of stimuli that it encounters. Classical conditioning is illustrated in Figure 3.1.

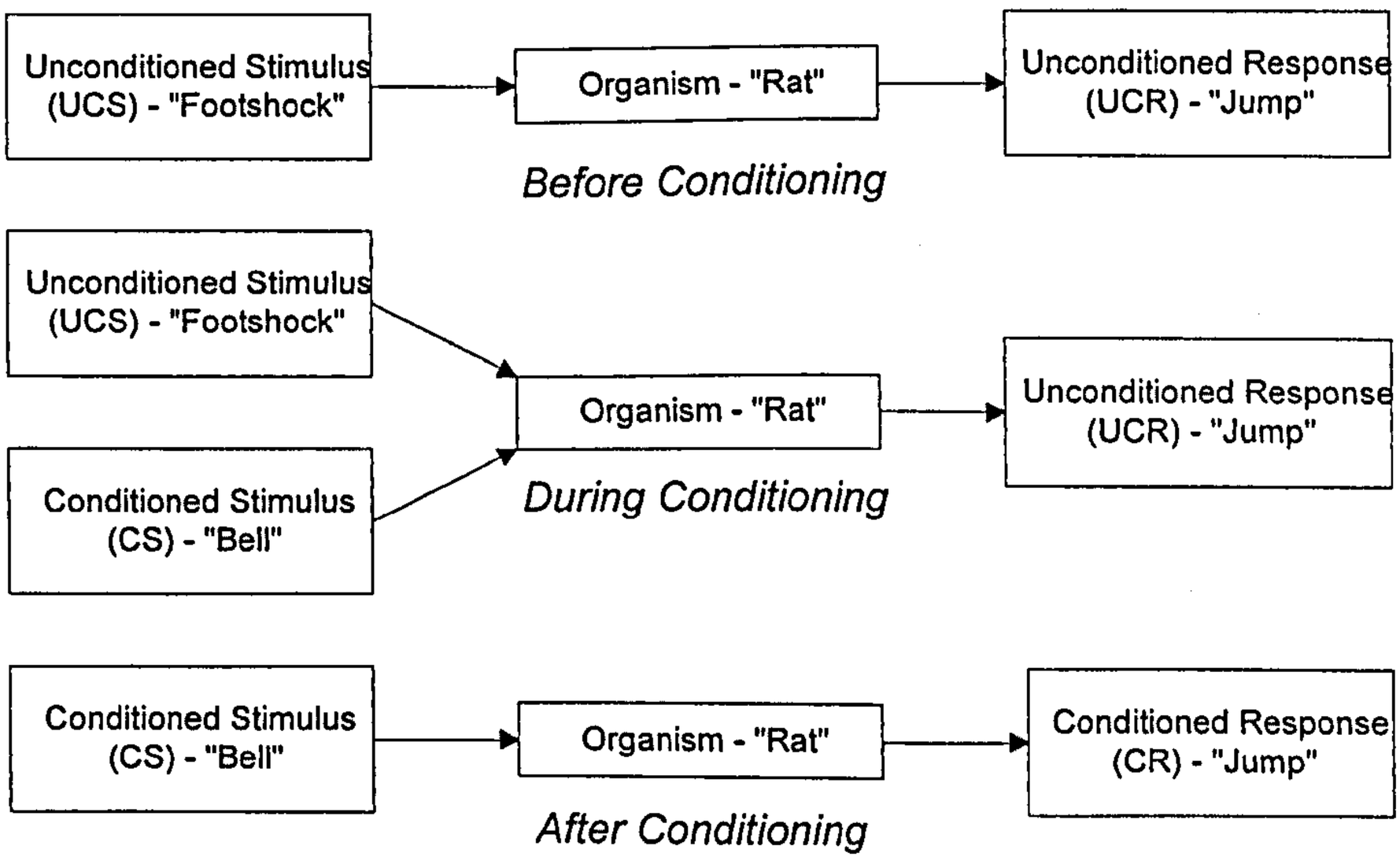


Figure 3.1 - Classical Conditioning

Classical conditioning assumes that an organism possesses knowledge about a basic set of reinforcing stimuli called 'primary' or unconditioned stimuli (UCS). These stimuli automatically trigger an unconditioned behavioural response (UCR) from the organism (such as freezing or jumping) because this response is beneficial to the survival of the organism. Classical conditioning requires that an organism is able to develop associations between the UCS and other 'secondary' or conditioned stimuli (CS). These CS only have reinforcing properties because of their association with the UCS, to the extent that the CS become able to elicit the required response in the absence of the UCS. For example, a rat (the organism) is given a foot shock (a negatively reinforcing UCS) a few seconds after the sound of a bell (the CS) is presented. The appropriate unconditioned response (UCR) is to jump in order to avoid the foot shock. The bell becomes associated with the footshock so that the rat learns to predict the footshock from the bell alone, and is thus able to produce the appropriate response, now a conditioned response (CR). The nature of the conditioning is often differentiated by referring to behaviour that is positively reinforced as *appetitive*, and behaviour that is negatively reinforced as *aversive*.

3.2.2 - Instrumental Conditioning

This type of conditioning (sometimes called operant conditioning) is different from classical conditioning in that behavioural responses are governed by the consequences of actions, sometimes referred to as the 'law of effect' [Schwartz & Robbins, 1995]. Conditioning is *instrumental* in that an organism must operate on its environment in order to receive a reinforcement, and the learning of this operant or instrumental stimulus (IS) comes to control behaviour. For example, a rat (the organism) learns to predict a food reward (a positively reinforcing UCS) by associating the UCS to the sight of a lever (the IS). Therefore each time the rat sees the lever, it will press the lever (a conditioned

response or CR) in order to receive a food reward. This type of conditioning is mostly studied in appetitive behavioural experiments. Instrumental conditioning is illustrated in Figure 3.2.

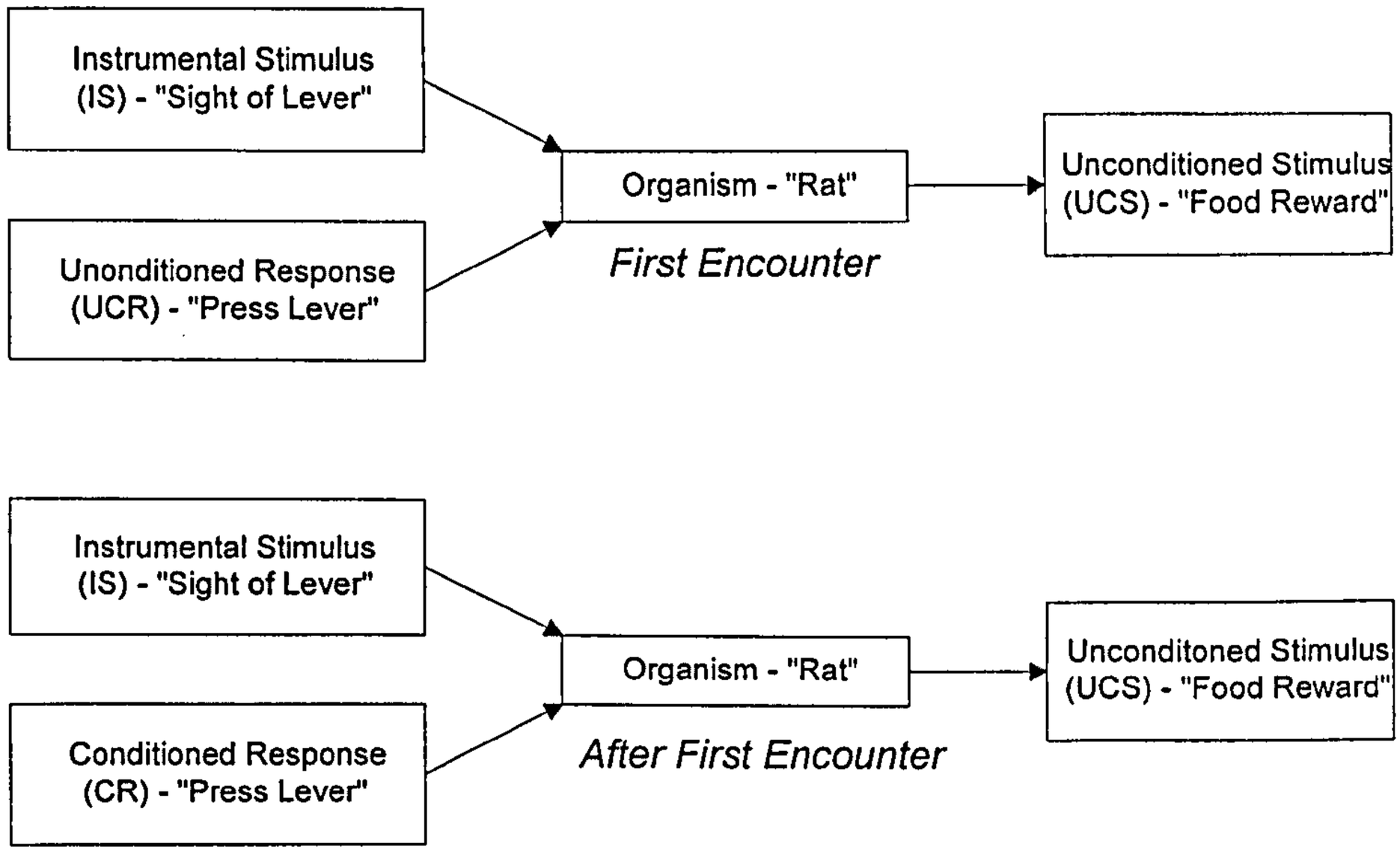


Figure 3.2 - Instrumental Conditioning

There are important differences between classical and instrumental conditioning. In classical conditioning, associations occur between stimuli so that the conditioned response (CR) is triggered as a result of the conditioning of the CS to the reinforcing UCS. With instrumental conditioning, however, the conditioned response actually controls the behaviour because the IS is conditioned to the appropriate motor action (CR) in order to receive the reinforcing UCS. This difference is important because the adaptive critic system described in Chapter Two (and in more detail in Chapter Five) is based on a model of reinforcement learning that enables classical but not instrumental conditioning [Barto et al. 1983]. Therefore, this adaptive critic design has not considered the full capabilities embodied in the theories of animal conditioning.

3.2.3 - Reinforcement and Reward

Reinforcing stimuli that depend on an instrumental action (e.g. lever pressing) have the effect of altering the future probability of repeating that action. The action itself may be any action from an entire repertoire of possible behavioural actions. Reinforcing stimuli can thus have two possible values: *positive reinforcers* have the effect of increasing the probability of an action, and *negative reinforcers* have the effect of decreasing the probability of an action. To complicate matters, these reinforcements can be omitted or terminated having the opposite effect of increasing or decreasing the probability of the action upon which they are dependent. Table 3.1 summarises the relationship between reinforcement and the probability of a behavioural action. The up and down arrows represent the respective increase or decrease in the probability of an action based on the presentation or withdrawal of a stimulus that could be either positively reinforcing or negatively reinforcing.

	Positive	**Negative**
Stimulus Presentation	↑	↓
Termination/ Omission	↓	↑

Table 3.1 - Change in Action Probabilities (Adapted from Gray [1991])

It can be seen that the change in action probability for the presentation of a positively reinforcing stimulus is the same as for the withdrawal of a negatively reinforcing stimulus, and that the change in action probability for the presentation of a negatively

reinforcing stimulus is the same as for the withdrawal of a positively reinforcing stimulus. Thus, any reinforcement learning system can function on the basis of a single reinforcement signal indicating the nature of the reinforcement. The lack of a positively reinforcing stimulus could be interpreted as a negatively reinforcing stimulus (or vice versa), which means that only one reinforcement signal is required [Bozarth, 1991]. The reinforcement signal forms the basis of a *reward system*. This system 'pulls' the animal towards a particular goal by using stimuli associated with an expected positive reward. Stimuli associated with an expected negative reward could equally be used to 'push' the animal away from non-goals. Therefore, a reward can be defined as a stimulus that increases or decreases the probability of any action with which it is regularly associated [Stein, 1980].

Although many researchers use the terms 'reward' and 'reinforcement' interchangeably, it has been argued that the two terms are different because 'reward' is more commonly used to represent a stimulus or event, whereas 'reinforcement' refers to the *process* of strengthening specific actions in accordance with the reward [Stellar & Stellar, 1985]. Rewards are only effective if their presentation is made contingent on the occurrence of a particular action, and this action in turn may have an effect on the environment (allowing for both classical and instrumental conditioning). Stimuli not associated with specific actions are therefore not rewards. This thesis will use 'reward' when referring to any stimulus that is reinforcing, and the term 'reinforcement contingency' to describe the dependency of a reward on specific behavioural actions. Changing the reinforcement contingency means to make the reward dependent on a different set of behavioural actions, and this can be seen to have the same effect as changing the reinforcement schedule described in Chapter Two.

3.2.4 - Biological Substrates of Reinforcement Learning and Relearning

Many researchers have attempted to explain the reward system in the brain in terms of the underlying neurophysiology, e.g. evidence from physiological and behavioural experiments such as lesion studies and brain stimulation. Gray [1991] proposes that the control of behaviour involves three separable subsystems based on rewards. The evidence to support *separable* subsystems comes from the way in which different lines of research have converged on the same set of brain structures. Most of this research comes from experiments with monkeys and rats, but is assumed to be equally applicable to humans. The control of behaviour is mediated through a number of operational (controlling) states in the brain, and these are elicited by particular reinforcing stimuli under the guidance of the subsystems [Gray, 1991]. These subsystems consider the relationship between the behaviours elicited by either the presentation, omission or termination of stimuli associated with rewards. The subsystems are the *Approach System*, the *Fight/Flight System*, and the *Behavioural Inhibition System*. These subsystems will be described in terms of their inputs, outputs, and the brain regions implicated in the functioning of each subsystem :-

1) The *Approach System* - a system that is responsive to stimuli associated with positive rewards, or the termination/omission of stimuli associated with negative rewards, as illustrated in Figure 3.3.

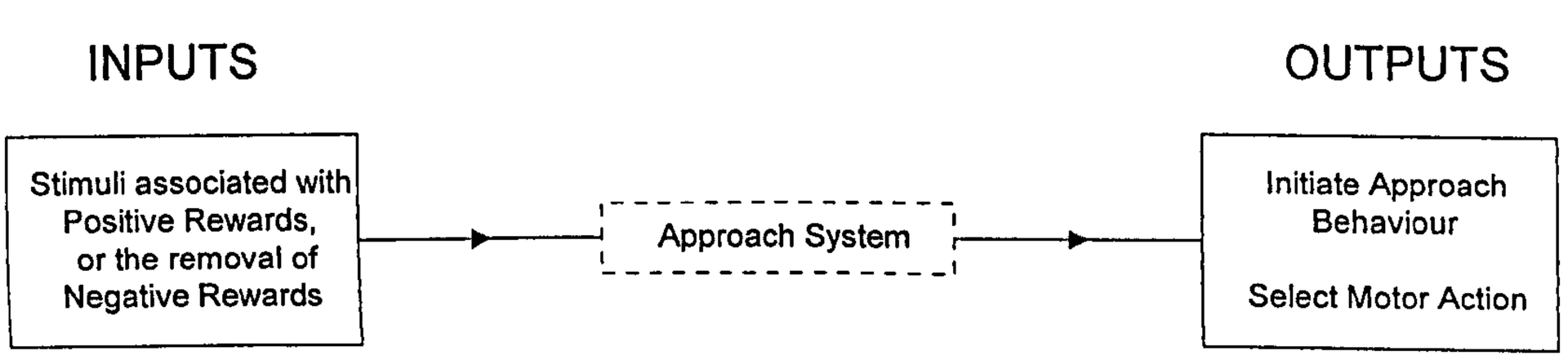


Figure 3.3 - The Approach System (Adapted from Gray [1991])

This system is responsible for approach behaviour, and may be considered to be the *positive reward system* of the brain. It allows an organism to respond to stimuli associated with positive rewards, or the removal of stimuli associated with negative rewards. For example, an organism will approach a food source (stimulus associated with positive reward), but only if it sees a predator leave the area (removal of stimulus associated with a negative reward). Such behaviour promotes the survival of the organism. Gray [1991] suggests that this system exercises a *selective* function in that it facilitates the selection of motor programs (a sequence of motor actions) for immediate execution, as well as playing a more general role in facilitating the performance of whichever motor program has already been selected. This system has implications for relearning because when reinforcement contingencies change, the relationship between stimuli and rewards also changes. It is this system that will be required to change the association between stimuli and their previously appropriate avoidance responses. Gray [1991] uses the evidence from lesion studies to suggest that this system is based around ascending projections from regions called the ventral tegmental area and substantia nigra in the brain stem. These projections innervate various regions of the basal ganglia, limbic system and neocortex using the neurotransmitter substance dopamine, and the importance of this neurotransmitter substance is discussed later in the chapter.

2) The *Fight/Flight System* - a system that is responsive to stimuli associated with unconditioned negative rewards, or the termination/omission of stimuli associated with positive rewards, as illustrated in Figure 3.4.

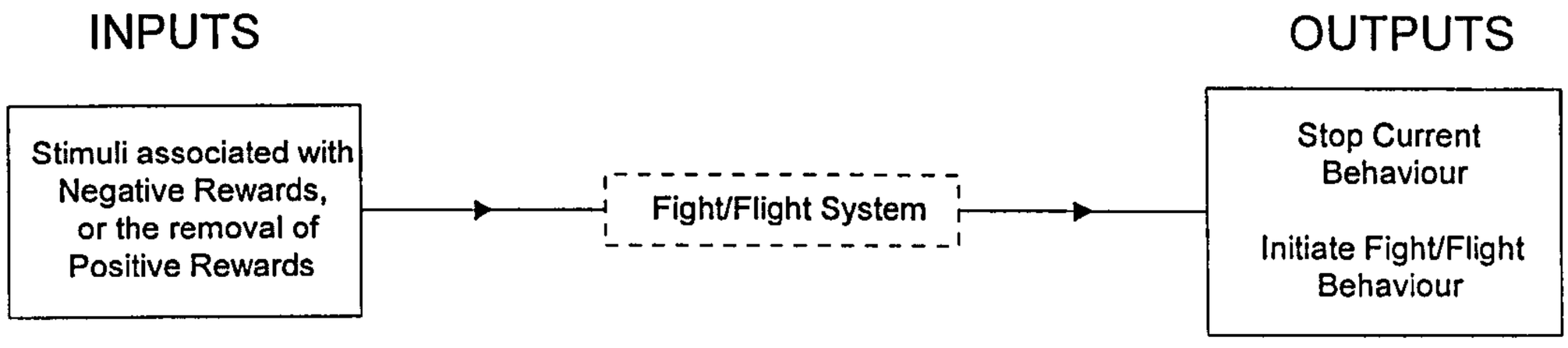


Figure 3.4 - The Fight/Flight System (Adapted from Gray [1991])

This system is responsible for initiating either confrontation or avoidance behaviours in response to stimuli associated with negative rewards, or the omission/termination of stimuli associated with positive rewards. For example, an organism may avoid a food source because of the threat of a predator (stimulus associated with negative reward), but may risk the presence of that predator because of the absence of food (omission of stimulus associated with positive reward). These behaviours are essential whenever the organism is under threat and will greatly affect its chances of survival. Gray [1991] argues that this system is essentially suppressed until it receives a signal to indicate that a fight or flight situation has been detected, and then uses a decision mechanism that selects between confrontation or avoidance behaviours. In terms of relearning and reinforcement contingencies, it is this system that will be required to rapidly initiate the appropriate behaviour selected by the decision mechanism whenever reinforcement schedules are changed. This effectively stops the current motor program and replaces it with some other behaviour, perhaps innate physiological responses such as freezing or increased heart rate. Gray [1991] suggests that this system is activated by a part of the brain called the amygdala based on evidence from experimental lesion studies. He also cites evidence to suggest that terminating the current motor program is the responsibility of a brain region called the septo-hippocampal system, and that the decision mechanism is located in a region called the hypothalamus.

3) The *Behavioural Inhibition System* - a system that is responsive to stimuli associated with negative rewards, the termination/omission of stimuli associated with positive rewards, or novel stimuli. This is illustrated in Figure 3.5.

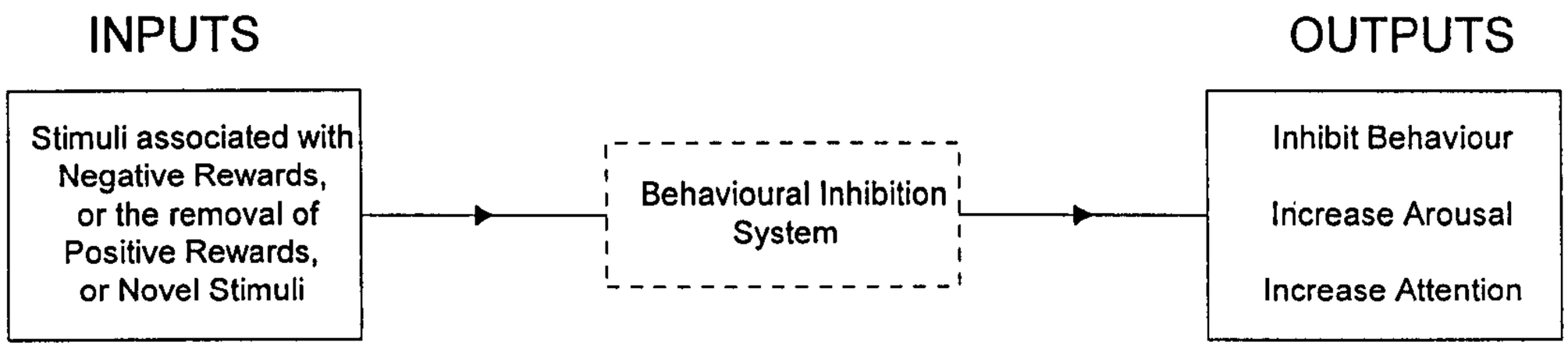


Figure 3.5 - The Behavioural Inhibition System (Taken from Gray [1991])

This system is responsible for interrupting whatever ongoing behaviour is occurring whenever the inputs shown in Figure 3.5 are received. This system is also responsible for incrementing the level of arousal so that the next behaviour is carried out with more speed or vigour, and increasing attention so that more information is taken in (particularly with regard to novel features in the environment). Any one of the inputs will cause the system to elicit all of the outputs. The neural structures that form the Behavioural Inhibition System are still under investigation, but Gray [1995] proposes that this involves a group of neural structures found in the "Limbic System", and the ascending projections that innervate these regions using the neurochemicals noradrenaline and serotonin. The structures that comprise the limbic system are still the subject of some debate, but Brodal [1992] suggests that these include the amygdala, the septo-hippocampal system and the associated "Papez Circuit" (anterior thalamic nuclei, cingulate cortex, mammillary bodies), and areas of the temporal lobe and frontal cortex.

It is interesting to note that both the Fight/Flight System and the Behavioural Inhibition System receive the same inputs, but the Behavioural Inhibition System also

receives "novel" stimuli as additional inputs. Gray [1991] states that novel stimuli are detected by comparing predicted and actual stimuli, and this is carried out in the subiculum which is part of the septo-hippocampal system. This aspect will be considered when the subicular comparator hypothesis [Gray, 1995] is discussed later in the chapter. It is also interesting that both Fight/Flight System and Behavioural Inhibition System involve both the septo-hippocampal system and the amygdala. The apparent overlap between the functions of the amygdala and the septo-hippocampal system has been the focus of much investigation, and some attempts have been made to differentiate between the functionality of these neural structures. McDonald & White [1993] looked at the relationship between stimulus inputs and behavioural responses, and used evidence from behavioural and lesion studies to identify three distinct systems for the learning of different types of association :-

- The *Hippocampal System* is thought to acquire information on the relationships among stimuli and how they relate to behaviour, i.e. the location of stimuli and the order in which they appear. This is sometimes referred to as a stimulus-response contingency because behavioural responses are related to a specific sequence of input stimuli. This is supported by evidence from rats on the win-shift task in a radial maze. This task consists of a radial maze in which each arm contains a food pellet, and rats must visit each arm only once without revisiting any arm in order to obtain a food pellet. The radial maze is in a room that has extra-maze cues, and rats would normally use these stimuli to differentiate between visited and unvisited arms of the maze to initiate the appropriate response (approach). Rats with lesions to the hippocampus were found to be impaired on this task, whereas lesions to the amygdala and dorsal striatum had no effect.

- The *Dorsal Striatum* is part of the basal ganglia, and is thought to mediate the formation of behaviours based on reinforced stimulus-response contingencies, i.e. where behavioural responses do not depend on the sequence or location of a set of input stimuli. This is supported by evidence from rats on the win-stay task in a radial maze. In this task, reinforcements (food pellets) are only available from arms of the maze that are lit. Rats are required to enter each lit arm twice per trial, thus receiving eight food pellets. Each pellet is given whenever the rat completes a stimulus (light) - response (approach) sequence, and this is independent of the spatial location of the arm because each trial involves a different set of lit arms. The food pellets are reinforcements because they strengthen the light-approach associations and weaken the dark-approach associations, essentially reinforcing the stimulus-response contingency. Rats with lesions to the dorsal striatum were impaired on this task, but lesions to the amygdala and hippocampus had no effect.

- The *Amygdala* system is thought to mediate the formation of behaviours based on the association of neutral stimuli with stimuli that have rewarding properties, i.e. the relationship between input stimuli and rewards. This is supported by evidence from rats on the conditioned cue preference task (CCP) in a radial maze. In this task, each arm of the maze may contain a food pellet, and rats are exposed to either a visual cue or a neutral cue in either the presence or absence of a food pellet. The rats are allowed to consume the pellet in the presence of either cue without making approach responses towards that cue. The rats must associate the visual cue with only the rewarding properties of the food (such as its smell or taste), or the reinforcing consequences of consuming the food. The food pellet can be differentiated by its smell or taste, but rats are not able to discern this difference until they are close to or have consumed the food.

McDonald & White [1993] describe this association as a 'stimulus-reward' association because only the rewarding properties of the food are contingent on the visual cue, and not the food itself. The visual cue thus acquires the ability to attract the rat even in the absence of a reinforcement, i.e. a food pellet. Rats with lesions to the lateral nucleus of the amygdala were impaired on this task, but lesions to the hippocampus and dorsal striatum had no effect.

3.3 - The Amygdala

The previous section introduced three different areas of the brain (amygdala, hippocampus and the dorsal striatum) thought to be involved in various aspects of reinforcement learning, and it is likely that these areas will also be involved in relearning processes. Some studies have focused on the relationship between stimulus inputs and behavioural responses, and particularly at how the hippocampus and septal regions are involved when this contingency is changed [e.g. Denham & McCabe, 1996]. If these regions are involved in detecting changes to input stimuli and their association with behavioural responses, they may also be involved in changing these responses as part of the relearning process. However, in order for appropriate behaviour to occur, there must also be some way in which changes to the rewarding properties of input stimuli can be considered. This means that any mismatch between expected and actual rewards needs to be detected and quickly acted upon. This is an appraisal process, and it seems likely from various lines of research that the reward aspect of this process is tied to the functioning of the amygdala [LeDoux, 1995], and its interaction with other brain regions involved in reinforcement (e.g. the basal ganglia). LeDoux [1989] argues that the amygdala is at the centre of this process, and predicts the rewarding properties of stimuli which allows the evaluation of input stimuli even in the absence of reinforcement. More recently, LeDoux

[1997] suggests that the connectivity of the amygdala enables it to influence attention, memory and perception primarily in situations where the organism is facing danger. The amygdala provides a fast neural pathway by which behavioural responses can be initiated in response to threatening or dangerous stimuli. This view of the amygdala focuses its involvement on negative reward learning, and is supported by work on fear conditioning [LeDoux, 1992; Maren & Fanselow, 1996], and the acquisition and retention of avoidance behaviours [Liang & McGaugh, 1983]. These studies have begun to uncover plausible neural circuits for the involvement of the amygdala in specific reinforcement learning and memory processes, and have identified the pathways that provide sensory inputs into the amygdala, the contribution of individual regions within the amygdala, and subsequent output pathways to other areas of the brain. The amygdala has been implicated in almost every experimental task used to study reward representation in the brain, yet progress towards understanding the functional organisation of the amygdala has only been achieved very recently [LeDoux, 1995]. The amygdala is not exclusively involved in reinforcement learning, and its other functions (i.e. as a mediator of arousal) have also been the subject of investigation [e.g. Gallagher & Holland, 1994; LeDoux, 1995]. The remainder of this chapter looks in more detail at the anatomy and function of the amygdala, and in particular at its neurochemical interaction with the basal ganglia structures involved in reinforcement learning.

3.3.1 - The Amygdaloid Complex

The amygdala is a subcortical region of the brain located in the anterior part of the temporal lobe. It is often referred to as the 'amygdaloid complex' because it consists of a considerable number of interconnected subnuclei which are highly connected to other parts of the brain [Sarter & Markowitsch, 1985]. The subdivision of the amygdala has always

been difficult, and so the amygdala is frequently identified as a single functional unit which hides the true nature of its structure and function. Where the amygdala has been subdivided, it usually consists of two main divisions: the *centromedial* division (central and medial amygdaloid nuclei) and the *basolateral* division (basal and lateral amygdaloid nuclei). Many researchers have referred to these divisions in their experimental studies, but have not agreed upon exactly which individual groups of subnuclei actually belong to which divisions. This has resulted in confusion and dispute regarding the validity of some of the results. Amaral et al. [1992] argue that the subdivision of the amygdaloid nuclei is still an ongoing process, and that it is the complexity of the amygdala coupled with the relatively poor anatomical attention it has received to date that has caused these problems. In an attempt to introduce some standardisation, they describe the intrinsic organisation of the primate amygdala in terms of three major subdivisions :-

- The *deep nuclei* - lateral, basolateral, basomedial and accessory basal
- The *superficial nuclei and areas* - anterior cortical nucleus, medial nucleus, nucleus of the lateral olfactory tract, periamygdaloid cortex, posterior nucleus
- The *central nucleus and remaining amygdaloid nuclei* - anterior amygdaloid area, intercalated nuclei, amygdalo-hippocampal area.

For the purposes of this chapter, discussion will focus on three of the main subnuclei groups described above: the *lateral*, *basolateral* and *central* nuclei. This follows the convention used by LeDoux [1990] and his work on relating the mechanisms of fear conditioning to individual amygdaloid nuclei in rats. It is assumed that the nuclei in rats more or less correspond to the nuclei identified in primates by Amaral et al. [1992], even though there may well be differences between these species.

3.3.2 - Afferent Connections to the Amygdaloid Complex

There are numerous brain regions that send information to nuclei in the amygdaloid complex. Carlson [1986] in a review of the anatomy of the amygdaloid complex states that this structure receives information from various sensory areas: the visual association cortex of the inferior temporal lobe, the auditory association cortex of the superior temporal lobe, the olfactory bulb, as well as from various thalamic and hypothalamic nuclei. The connections to the amygdaloid complex allow it to receive a great deal of sensory information from many different sensory modalities. Most of this information goes to the lateral nucleus of the amygdala [Bordi & LeDoux, 1992]. The lateral nucleus is not the only amygdaloid nucleus to receive afferent connections. Rolls [1990] reports that the basolateral nucleus receives projections from the ventral tegmental area, and he also describes reciprocal projections between the basolateral nucleus and the orbital prefrontal cortex. LeDoux [1997] suggests that the connections between the orbital prefrontal cortex and the basolateral amygdala are particularly involved in storing memories about the rewarding properties of stimuli as they are experienced. The significance of the orbital prefrontal cortex will be discussed in relation to how the amygdaloid complex keeps a record of rewards as they are actually encountered.

3.3.3 - Efferent Connections from the Amygdaloid Complex

In addition to receiving information, the amygdaloid complex sends information to a number of brain regions. Carlson [1986] states that the amygdaloid complex sends projections to the cortex, basal forebrain, hypothalamus, dorsomedial thalamus, and various basal ganglia and brain stem nuclei. This allows it to influence various reinforcement mechanisms and activate species-typical behaviours. The interest in these behaviours lies in the fact that they are elicited from the central nucleus of the amygdala,

the origin of most of the efferent projections from the amygdaloid complex. Kapp et al. [1984] looked closely at the role of the central nucleus in species-typical responses to threatening stimuli in rabbits. They cited a number of studies that found that conditioned fear responses (e.g. freezing) were affected by lesions of the central amygdaloid nucleus, and suggested that this nucleus works in conjunction with forebrain and brain stem structures to enable various aspects of aversive conditioning. LeDoux et al. [1988] showed that projections from the central nucleus to the lateral hypothalamus mediated autonomic aspects of conditioned fear, and projections from the central nucleus to the central gray (an area in the brain stem) mediated behavioural aspects of conditioned fear. This latter finding is of significance because if the central nucleus influences the control of learned behaviours, then it may also be involved when the relearning of new behaviours is required. This could be in terms of changing the associations between confrontation/avoidance behaviours and stimulus inputs, consistent with the idea of the fight/flight system presented by Gray [1991].

The connections from the amygdala to the cortex are also important. LeDoux [1997] notes that the projections from the cortex to the amygdala are considerably fewer than the projections from the amygdala to the cortex. He also notes that the amygdala projects back to cortical sensory processing areas from which it does not receive inputs, which enables the amygdala to exert a direct influence on the cortex. The nature of this influence will be discussed later in relation to the neurotransmitter glutamate.

3.4 - The Basal Ganglia

The basal ganglia are a system of subcortical structures that lie beneath the neocortex and surround the thalamus, including the caudate nucleus, putamen and globus

pallidus, as well as the subthalamic nucleus and the substantia nigra [Carlson, 1986]. Consistent with their anatomical location, the basal ganglia receive topographical projections from almost the entire neocortex and parts of the limbic system, and project to frontal areas of the cortex (via thalamo-cortical projections) and the reticular formation. It has been suggested that the basal ganglia are organised into a number of largely separate circuits which appear to unite cortical and thalamic regions dedicated to performing a common function [Alexander & Crutcher, 1990]. Some of these circuits have been functionally identified, e.g. the "motor" circuit centred on the supplementary motor area and motor cortex, and the "oculomotor" circuit centred on the frontal eye fields. Other circuits are not so obviously tied to motor functions, e.g. the "limbic", "orbitofrontal" and "dorsal prefrontal" circuits. Jackson & Houghton [1995] note that the organisation of these circuits seems to follow a similar pattern, suggesting that their computational functions may be equivalent. The basal ganglia will now be described in terms of their afferent and efferent connections.

3.4.1 - Afferent Connections to the Basal Ganglia

The striatum is located in the anterior part of the basal forebrain, and can be considered the 'input structure' of the basal ganglia because it receives projections from the cortex, thalamus, midbrain and limbic structures. It is composed of two main segments: the *dorsal striatum* consisting of the caudate nucleus and putamen, and the *ventral striatum* consisting of the nucleus accumbens. Groenewegen et al. [1991] state that the cellular design of the ventral and dorsal striatum is comparable, but the two striatal regions differ in terms of their innervation by the neurotransmitter dopamine. The ventral striatum is innervated by dopamine projections from the ventral tegmental area (a nucleus in the ventral tegmentum of the midbrain), and the dorsal striatum is innervated by dopamine

projections from the substantia nigra (a darkly stained region also in the tegmentum). Both the dorsal and ventral striatum project to the globus pallidus with inhibitory projections, and this inhibition is via two distinct pathways (dorsal and ventral) that use the neurotransmitter substance GABA. Under normal conditions these two pathways are sensitively balanced, but it has been suggested that they can be chemically modulated by the neurotransmitter dopamine [Jackson & Houghton, 1995]. The ventral striatum also has GABA projections to the substantia nigra pars compacta, the origin of dopamine projections to the dorsal striatum. The substantia nigra is actually part of the tegmentum, but it is usually included in the basal ganglia because of its interconnection with the striatum and globus pallidus [Carlson, 1986]. The significance of the substantia nigra will be discussed later in terms of dopamine.

3.4.2 - Efferent Connections from the Basal Ganglia

The globus pallidus may be considered the 'output structure' of the basal ganglia because it projects to the thalamus, which in turn projects to a number of motor output and related cortical areas. The ventral pallidum receives projections from the ventral striatum and projects to the mediodorsal thalamus, and the dorsal pallidum receives projections from the dorsal striatum, and projects to the medial and lateral thalamic nuclei [Brooks, 1986]. These pathways activate motor actions that are represented as motor programs. These programs are complex sequences of cortical information, and a description of them is beyond the scope of this work. It is sufficient to say that these motor programs could represent species-typical behaviours (amongst other types of behaviours) that would be initiated as a result of detecting changes to reinforcement contingencies.

3.5 - Interaction of the Limbic System and Basal Ganglia

The previous sections have described structures in the limbic system (primarily the septo-hippocampal system, amygdaloid complex and associated cortical areas) and the motor system (ventral and dorsal striatum, substantia nigra and globus pallidus). A number of researchers have proposed theories to explain the interaction between the limbic system and motor system, such as the limbic comparator hypothesis [Brooks, 1986] and the subicular comparator hypothesis [Gray, 1995]. These hypotheses have strongly influenced the model proposed later in this chapter, and are therefore described here.

3.5.1 - The Limbic Comparator

Brooks [1986] proposed the *limbic comparator* hypothesis, suggesting that interactions between the limbic system and cortical areas related to motor actions are essential for learning what to do in a motor task (motor actions) and how to do it best (motor skill). This hypothesis was prompted by studies of motor learning which showed that monkeys develop motor skills after having learned the motor actions, such that inappropriate sequences of motor actions produce 'error' signals in the anterior cingulate cortex. The limbic system used by Brooks in his hypothesis was defined as a functional entity made up of structures that receive inputs from the hypothalamus and midbrain, and project to the thalamus and various cortical areas. This therefore includes the amygdala and the septo-hippocampal system. Brooks argues that the interaction of the limbic system with the basal ganglia can lead to the effects of comparator action being carried to the ventral striatum and dorsal striatum. Figure 3.6 illustrates the interactions of interest.

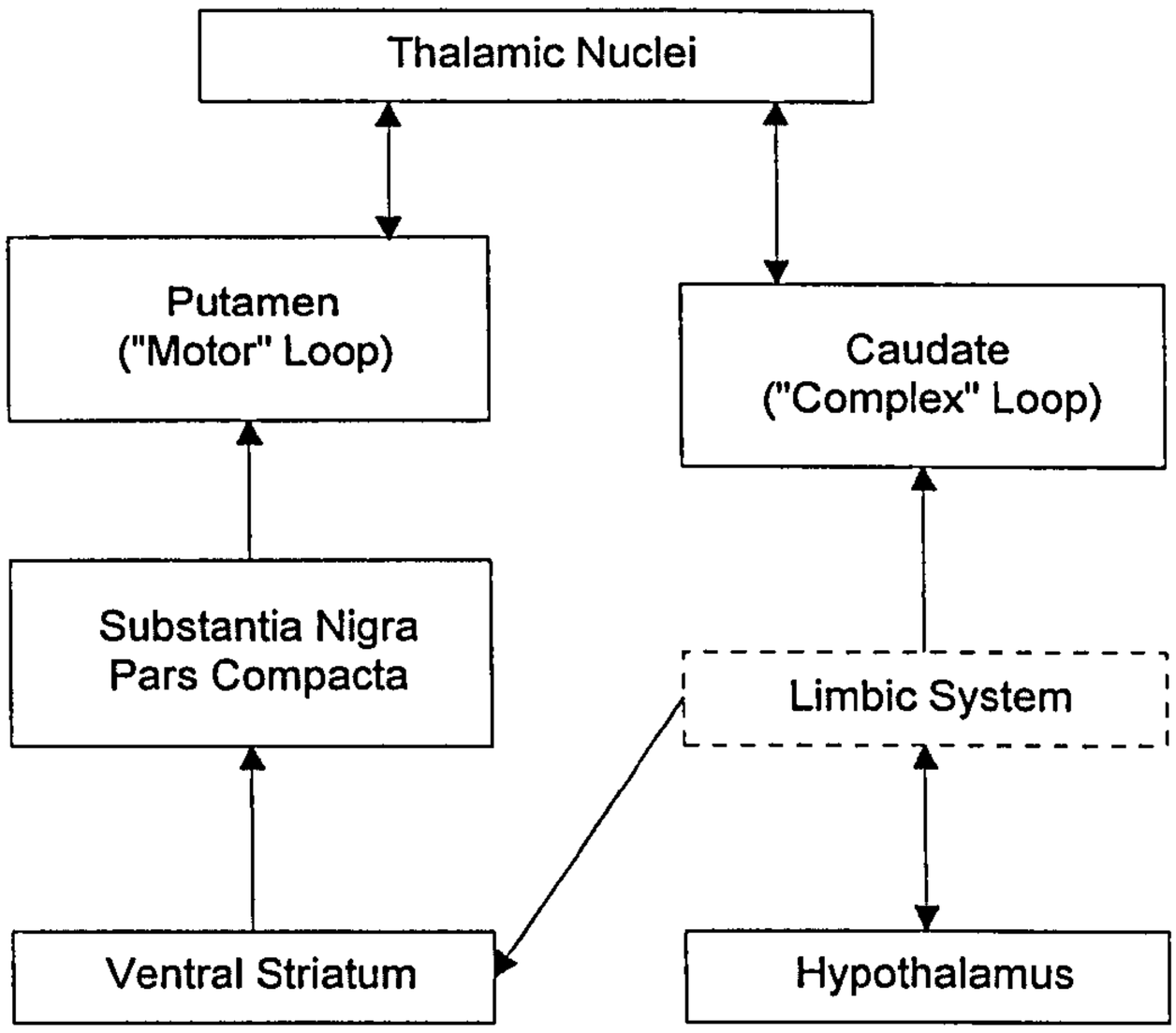


Figure 3.6 - Limbic-Motor Interactions (Adapted from Brooks [1986])

The diagram shows that the limbic system receives signals from the hypothalamus, which Brooks [1986] suggests is the origin of information about biological drives e.g. hunger, thirst etc. The limbic system influences two loops located in the caudate nucleus and the putamen. Each loop is composed of several parallel channels reached by limbic projections. These loops are not fully illustrated in Figure 3.6 because they would detail the various thalamic projections to and from cortical areas, and would overcomplicate the diagram. The "motor" loop is reached by connections from the limbic system to the ventral striatum, then through the substantia nigra pars compacta to the putamen. Brooks suggests that this loop is responsible for the learning of motor skills through connections via the thalamus to supplementary and primary motor areas. The "complex" loop is reached by connections from the limbic system to the caudate nucleus. Brooks suggests that this loop is responsible for controlling and regulating the assembly and management of behavioural actions in appropriate combinations through connections via the thalamus to higher association areas of the cortex. The limbic system can therefore influence both motor

actions and motor skills through its interaction with the basal ganglia. Brooks [1986] argues that the most important limbic structure for encoding the relevance of actions in the context of task and environment is the amygdaloid complex because it denotes the importance of an event by providing an 'affective bias'. He supports this argument by using evidence from lesion studies on monkeys indicating that the amygdaloid complex enhances behavioural stability appropriate to the environment, and also enhances the learning of the importance of visual and other sensory cues. These capabilities are clearly essential for social behaviour and the survival of animals in their natural habitat. An affective bias produced by limbic signals is thus also important for motivation because this bias reinforces the selective actions of higher-level cortical areas, which have been implicated in intelligent behaviour.

3.5.2 - The Subicular Comparator

Gray [1991] suggested that a comparator mechanism is an essential part of what he called the 'Behavioural Inhibition System', described earlier in this chapter. This was formalised as the *subicular comparator* hypothesis because the heart of the comparator function was attributed to the subicular area of the hippocampal formation [Gray, 1995]. The operations performed by the subicular comparator can be summarised as a sequence of steps as follows :-

1) Receive sensory information describing the current (perceived) state of the world
2) Obtain information regarding the current motor program
3) Compare perceived stimuli with past regularities stored in memory
4) Obtain information stored in memory that describes past regularities relating stimuli to behavioural responses

5) Use this information to predict the next expected state of the world

6) Compare the predicted with the actual next state of the world

7) Decide if there is a match or mismatch between the predicted and actual state of the world

8) If there is a match, do steps 1) to 7) again, otherwise go to step 9)

9) Bring the current motor program to a halt

10) Decide if the predicted state of the world is associated with a negative reward, if yes then go to step 11)

11) Obtain further information to resolve the difficulty that has interrupted the current motor program.

The steps described above are performed primarily by the septo-hippocampal system and the Papez Circuit, with comparisons performed in the subiculum. Steps 1 to 8 look at the relationship between stimuli and responses, and step 9 essentially provides a stop signal to inhibit the current behaviour. It is not until step 10 that the relationship between stimuli and rewards is actually considered. The subicular comparator hypothesis therefore seems to suggest that one way to activate relearning processes is to detect changes to perceptual stimuli (e.g. novel stimuli), or changes to the association between perceptual stimuli and behavioural responses. The relationship between stimuli and rewards is more consistent with the limbic comparator hypothesis (involving the amygdaloid complex), and is another way by which relearning processes could be activated.

3.6 - Interactions Between Neurochemical Systems

The previous sections looked at the amygdaloid complex and the basal ganglia, and outlined the main connections to and from these neural structures. Communication between neural structures is achieved by means of chemical signals between neurons in a variety of brain regions, and this is enabled by the interaction of several neurochemical substances. These substances have been implicated in mechanisms of reinforcement learning and relearning in the brain. This section will describe the involvement of the main neurotransmitters by identifying distinct neurochemical systems, and describing the effect these systems have on learning and relearning. Table 3.2 lists some of the neurotransmitter substances, and their hypothesised effects according to Carlson [1986] :

Neurotransmitter	Hypothesised Effect
Dopamine (DA)	Inhibitory
Noradrenaline (NA)	Inhibitory
Serotonin (5-HT)	Inhibitory
Acetylcholine (ACh)	Excitatory
Gamma-Aminobutyric Acid (GABA)	Inhibitory
Glutamate (Glu)	Excitatory

Table 3.2 - Neurotransmitter Substances (Adapted from [Carlson, 1986])

In Table 3.2, the hypothesised effects are given as either *inhibitory* or *excitatory*. An excitatory neurotransmitter will increase the propensity of a neuron to fire, and an inhibitory neurotransmitter will decrease the propensity of a neuron to fire. A 'neurotransmitter' should be distinguished from a 'neuromodulator', which is a

neurochemical substance that acts with secondary messenger systems to directly or indirectly modify the effects brought about by other neurotransmitter substances, but does not cause these effects by itself [Mogenson & Yim, 1991]. This distinction means that a neuromodulator can be either inhibitory or excitatory depending on the neurochemicals with which it interacts.

Hestenes [1992] uses the classification reported by Strange [1988] whereby neurotransmitters fall into two general categories according to their rate of receptor response: fast receptors and slow receptors. Fast receptors have a response rate on the order of milliseconds, whereas slow receptors respond on the order of hundreds of milliseconds or slower. Hestenes argues that only the fast receptors are fast enough to deal with the observed rate of information processing in the brain, and he therefore concludes that slow receptors perform a modulatory function. The fast neurotransmitters are either excitatory (e.g. glutamate) or inhibitory (e.g. GABA), but the slow receptors are not so easily classified because they display a variety of modulatory effects. The slow neurotransmitters include the monoamines (e.g. dopamine, noradrenaline, serotonin, acetylcholine) that originate in a few groups of nuclei in the brain stem and midbrain [Hestenes, 1992]. These neurotransmitters will now be discussed in more detail.

3.6.1 - The Dopaminergic System

Dopamine (DA) has been implicated in several important functions, including switching between motor actions, modulation of attention mechanisms, and reward learning. Robbins [1992] outlines three main dopaminergic systems in the brain :-

- the *mesolimbic* DA system projecting from the ventral tegmental area (VTA) to the ventral striatum

- the *mesocortical* DA system also projecting from the VTA to the prefrontal cortex, anterior cingulate cortex, entorhinal cortex, and olfactory bulb
- the *mesostriatal* DA system projecting from the substantia nigra pars compacta to the dorsal striatum.

Dopamine is also found in the thalamus, locus coeruleus, cerebellum, hypothalamus, median eminence, pituitary gland and certain sensory systems, but the implications of this will not be discussed here. The three main DA systems are of interest because of the different role that dopamine appears to play in each system. Table 3.2 indicates that DA is an inhibitory neurotransmitter, but there is still no agreement as to whether DA is really inhibitory or excitatory. Mogenson & Yim [1991] suggest that DA is actually a neuromodulator because it can be either excitatory or inhibitory depending on the neurochemicals with which it interacts. However, they cite some empirical evidence showing that DA alone can produce direct postsynaptic effects, so it cannot be strictly classified as a neuromodulator. Mogenson & Yim [1991] therefore suggest that DA only has a neuromodulatory effect in the mesolimbic DA system (which influences the 'throughput' to the ventral striatum), and exerts a selective influence on limbic inputs in order to focus the integration of signals from limbic to motor systems. Dopamine elsewhere is inhibitory, in accordance with the hypothesised effect of this neurotransmitter. They suggest that the neuromodulatory effects in the mesolimbic DA system could be due to the presence of the neuropeptide cholecystokinin (CCK), which is found in relatively high concentrations in both the nucleus accumbens and the mesolimbic DA system.

The mesolimbic DA system was considered by Koob [1992] who suggested that this system (and only this system) is involved in activating the effects of rewards, because the results from experiments with rodents show that rewards produce a behavioural

activation that can only be disrupted by lesions to the mesolimbic DA system. This is supported by experiments involving selective destruction of the DA projection to the nucleus accumbens, which decreased the locomotor activity normally induced by novel environments, and the motor activity normally induced by food presentation [Robbins & Everitt, 1992]. Koob [1992] has reviewed a number of other studies that suggest that the mesolimbic DA system is involved in attention mechanisms. He reports that these studies show disruption of the mesolimbic DA system leads to problems with normal attention mechanisms, e.g. causing behavioural perseveration, an inability to ignore distractions caused by irrelevant information, a decrease in behavioural switching and flexibility, and a difficulty in reversing previously learned behaviours. The hypothesis proposed by Koob is that the mesolimbic DA system modulates a filtering and gating mechanism in the nucleus accumbens for signals received from limbic structures such as the amygdala, hippocampus and frontal cortex. These signals carry information about basic biological drives and motivation, and are ultimately turned into motor acts by the output circuitry of globus pallidus. This is supported by Mogenson & Yim [1991], who quote Willner [1983]:

"... biologically significant stimuli converge on the VTA and influence the firing of the mesolimbic DA system; activity in this system modulates the transfer of information through the nucleus accumbens, which acts as a 'limbic-motor interface', receiving inputs from the amygdala and other structures traditionally implicated in emotional and motivational behaviours, and sending its output to the motor system."

The idea that mesolimbic DA provides a switching mechanism in the nucleus accumbens is supported by other researchers. Hestenes [1992] suggests that the nucleus accumbens is a gate through which the limbic system exerts control over behavioural

output from the basal ganglia, such that DA functions as a gain control parameter that opens the nucleus accumbens gate when sufficiently present, and closes it when sufficiently absent. He argues that the projection from the VTA to the nucleus accumbens is an execution pathway, and the projection from the substantia nigra to the dorsal striatum is a selection pathway. The GABAergic projection from the nucleus accumbens to the substantia nigra thus provides a means of co-ordinating the gains, and thus influences the output through both pathways [Hestenes, 1992]. The idea that dopamine functions as a gain control parameter will be revisited in Chapter Four with reference to neuromodulation.

The switching hypothesis was also supported by Weiner [1990] who investigated latent inhibition (LI), which may be defined as a decrement in conditioning to a stimulus as a result of preexposure to that stimulus without reinforcement. Latent inhibition allows for stimulus selectivity because there is a bias towards potentially important stimuli, and this is due to the devaluation of stimuli regarded as inconsequential in the past. This phenomenon is considered to be an important reflection of attentional processes, and has been extensively studied by neuroscientists interested in the neural substrates of attention. The LI paradigm involves exposing an animal to a stimulus under two opposite reinforcement contingencies. It may be noted that this is equivalent to changing the reinforcement schedule as discussed in Chapter Two. In the preexposure stage, the stimulus is not followed by a reward and is considered irrelevant. In the conditioning stage, the stimulus is followed by a reward and therefore becomes relevant. Thus, during the conditioning stage, the same stimulus carries conflicting signals of relevance and irrelevance. To exhibit LI, the animal must continue to respond to the stimulus as irrelevant even though it comes to signal a reward, and animals are normally under the control of their previous learning of irrelevance rather than the new, changed reinforcement contingency. Weiner [1990] cites empirical evidence to suggest that disruption of the mesolimbic DA system interferes with

LI such that there is a rapid 'switching' of responding upon the introduction of the reward in the conditioning stage. Disruption of other DA systems has no effect on LI. Weiner therefore suggests that switching is mediated by the ventral striatum, such that increased DA in the nucleus accumbens promotes switching, and decreased DA in the nucleus accumbens eliminates switching and gives rise to perseverative behaviour. In contrast, decreased DA in the caudate gives rise to increased switching. These data suggest that the dorsal striatum is responsible for the continued execution of behavioural sequences, and that the ventral striatum is only responsible for switching between different or conflicting behavioural sequences. This is important to relearning because when a change in reinforcement contingencies is detected, there should be an expected increase in switching between responses. This suggests that relearning will cause an increase of DA in the ventral striatum, and a decrease of DA in the dorsal striatum.

3.6.2 - The Noradrenergic System

The noradrenergic system originates in the locus coeruleus and lateral tegmental group, and projects through the dorsal noradrenergic bundle to the septal nuclei, hippocampal formation and the amygdala [Robbins et al., 1985]. Noradrenaline is thought to modulate selective attention and long-term memory storage by pattern enhancement [Hestenes, 1992]. The involvement of noradrenaline in attentional processing is suggested because the locus coeruleus (LC) also projects to widespread areas of the cortex, and this anatomical arrangement is more consistent with a general modulatory influence than the conveyance of specific information [Bunsey & Strupp, 1995]. Table 3.2 shows that noradrenaline (NA) is inhibitory. Segal [1985] argues that NA of LC origin can suppress the spontaneous activity of many neurons in the brain and therefore focus their reactivity to certain stimuli. This is part of the *Selective Attention* hypothesis, supported by Robbins et

al. [1985] who show that the firing rates of neurons in the LC are highest when an animal is most vigilant to environmental events (e.g. when orienting to a startling stimulus) but lowest when the animal is least attentive to the external environment (e.g. during eating or sleeping). Posner & Peterson [1990] describe three functionally distinct neural systems believed to be involved in the regulation of selective attention :-

- an *anterior* attention system (anterior cingulate and supplementary motor areas) related to volitional control and awareness
- a *posterior* attention system (posterior parietal cortex, pulvinar and superior colliculus) which controls spatial orientation
- a *vigilance* system (locus coeruleus) which functions to place the anterior and posterior systems into an alert state, thereby enhancing attentional processing in these systems.

The LC is therefore implicated in attentional mechanisms when vigilance is required, such that noradrenaline increases the signal/noise ratio of signals arriving at a number of brain regions [Oades, 1985]. The LC has reciprocal connections with the raphe dorsalis (RD), the origin of inhibitory serotonergic projections to both the nucleus accumbens and the dorsal striatum, and Hestenes [1992] describes how this causes RD output to track how the LC responds to significant stimuli. Indeed, there is evidence to suggest that the regulation of noradrenergic activity in the LC is under the control of serotonergic afferents [McCrae-Degueurce et al., 1985]. This observation is of passing significance given that serotonin is inhibitory. The projection from the RD to both ventral and dorsal striatum may serve to inhibit DA output, and Hestenes [1992] describes a mechanism by which the ventral and dorsal striatum may be disinhibited by the LC (through the RD) in preparation for a more vigorous response. He suggests that this

mechanism could explain why lesions of RD projections to the ventral and dorsal striatum eliminate DA responses to certain stimuli. The connection from the RD to the LC could be a means by which neurons in the LC are accessed by other brain regions connected to the RD. This needs to be investigated, and may lead to an explanation of how and when the LC is aware of the need for vigilance. The RD does in fact also project to the septo-hippocampal system, and Weiner [1990] suggests that this provides an interrupt or 'stop' signal in response to negatively or non-reinforced stimuli. This therefore suggests a possible role for serotonin in behavioural inhibition, and a link between the LC and the septo-hippocampal system.

Sara [1985] suggests that LC neurons not only fire during situations requiring vigilance, but also in response to specific stimuli that have a biological significance by virtue of their previous association with reinforcement. This is useful in developing a model to explain how an animal is able to detect changes in reinforcement in the environment. More recently, Sara et al. [1995] investigated the effect of pharmacological manipulation of the noradrenergic system, and suggested that the LC is responsible for responding to changes in the environment and detecting novelty. Using recordings from single neurons in the noradrenergic nucleus of the LC, they found that these cells fire in burst when novel objects are first encountered by rats exploring the hole-board apparatus. It was observed that rats spend significantly more time investigating holes containing objects than empty holes in this apparatus. They then blocked the activation of the LC (using clonidine) which inhibited the release of NA within the LC and at noradrenergic terminals. This inhibition abolished the recognition of novelty in the rats, as shown by the elimination of a preference for holes with objects in them, but having no effect on the total time rats spend investigating holes. Sara et al. [1995] thus postulate that LC neurons respond robustly to the presentation of novel or significant stimuli enhancing attention to

these stimuli by the release of NA into target forebrain and other areas, which includes the amygdala. Information about stimuli that may or may not have been encountered before is therefore available to the amygdala at the same time as information regarding rewards. A thorough review of the underlying physiology of the LC and the mechanisms of selective attention is not provided in this work, but the NA projection from the LC to the amygdala is of interest. The LC could provide the amygdala with information that enables it to detect when the properties of stimuli have changed, and may be part of the process by which relearning mechanisms are activated. This can be related to reinforcement contingencies by evidence provided by [Sara, 1988] which showed experimentally that LC neurons during classical conditioning increase their firing rate whenever a stimulus-response contingency is first presented, but decrease their firing rate once the response is expressed behaviourally. Thus when novelty is detected in terms of a change in the reinforcement contingency, the LC may be able to signal this change to the amygdala. The importance of this signal will be considered when the neurotransmitter glutamate is discussed.

3.6.3 - The Cholinergic System

Gallagher & Holland [1994] postulate that attention is not mediated by a single system, but by several attentional systems consistent with the selective attention hypothesis. They suggest that the central nucleus of the amygdala regulates various attention mechanisms when important stimuli are first noticed or altered. Therefore, damage to this system interferes with the normal attention mechanisms engaged when an expectation about the occurrence of stimuli is violated [Holland & Gallagher, 1993]. This is consistent with the work of Hasselmo [1994], who suggests that acetylcholine is primarily released during the learning of new or unexpected stimuli, and not during the recall of previously learned information. Acetylcholine is excitatory, and innervates a

number of cortical systems in the basal forebrain, hypothalamus and brain stem. The basal forebrain influences various perceptual systems in the cortex, the hypothalamus activates autonomic responses, and the brain stem directly influences motor systems. The role of acetylcholine is also considered by Introini-Collison et al. [1996], who suggest that memory storage is regulated by acetylcholine release from the central nucleus of the amygdala. This is supported by the work of Kapp et al. [1994], who demonstrate that stimulation of the central amygdaloid nucleus leads to cortical arousal. It is likely that inputs from the central amygdaloid nucleus to the cortex indicate when stimuli are novel or unexpected, thereby influencing memory consolidation, and perhaps explaining why unexpected events are more strongly consolidated into memory.

3.6.4 - The Glutamatergic System

Glutamate is the primary excitatory neurotransmitter in the brain [Hestenes, 1992], released by some neurons to excite other neurons to a higher level of activity [Barinaga, 1990]. It has been implicated in Huntingdon's disease, a neurodegenerative condition where an abnormal glutamate metabolism is thought to be responsible for the death of large numbers of neurons, and it is known that a build-up of glutamate causes brain damage due to stroke, trauma and seizure. Glutamate projections to the ventral striatum have been identified from the hippocampus, amygdala, cingulate gyrus and insular cortex [Swerdlow & Koob, 1987]. These projections are of interest because it has been suggested that the mechanisms of switching in the nucleus accumbens are under the control of the glutamate projection from the subiculum to the ventral striatum [Cador et al., 1989]. The main excitatory chemical in the hippocampus is glutamate, which is also thought to play a role in the mechanisms of long-term potentiation and memory formation [Carlson, 1986]. Therefore, the subicular glutamate projection provides a pathway through which the

hippocampus can affect striatal processing, and is consistent with the subicular comparator hypothesis described by Gray [1995]. The glutamate projection from the basolateral amygdaloid nucleus to the nucleus accumbens is also of significance, because it provides another pathway through which the switching mechanisms in the nucleus accumbens can be influenced. The difference in the information provided by these two glutamatergic projections will be discussed later in the chapter.

Groves et al. [1995] cite evidence suggesting that glutamate release results in greater activation of glutamate receptors on dopamine terminals, and thus increases the release of dopamine. This is important because the interaction between glutamate and dopamine may have an influence on the functions normally attributed solely to the dopaminergic system described earlier. Glutamate projections should therefore be considered in any conceptual model that would account for reward learning and attention mechanisms in the brain.

3.7 - A Biological Model of the Amygdaloid Complex and Relearning

This chapter has discussed the widely accepted idea that the limbic system is involved in the learning and control of motor behaviour in response to reinforcement signals. It is now possible to develop a conceptual model to explain how limbic and basal ganglia structures are involved in the biological processes of learning and relearning, based on the evidence presented in the previous sections. The influence of the amygdaloid complex in activating specific biological systems during the processes of learning and relearning (such as perceptual, reward and motor systems) is central to this model, and is emphasised in terms of the involvement of specific amygdaloid nuclei and their connections to the neural centres that activate these systems. The model is based on the

hypothesis that the amygdaloid complex is involved in learning about rewards, and can detect when an expected reward is not received thus indicating that relearning is required. The amygdaloid complex is therefore able to *affect* a number of structures when relearning is required, which could be a means of focusing attention on these structures. The model draws upon elements from the subicular comparator hypothesis [Gray, 1995], limbic comparator [Brooks, 1986], the switching model (e.g. Weiner [1990]) and the selective attention hypothesis (e.g. Robbins et al. [1985]). The model is divided into three sections based on the main amygdaloid nuclei, each of which has a distinct role to play in the processes of reinforcement learning and relearning. A number of diagrams will be used to help explain the model. In each diagram, dotted arrows indicate excitatory connections, and labelled solid arrows indicate inhibitory connections. Thinner solid arrows are used to show other non-specific connections relevant to the model, and these connections are non-specific because they may involve pathways of different neurotransmitters and are left unlabelled.

3.7.1 - The Lateral Amygdaloid Nucleus as an Interface to Sensory Systems

The lateral amygdaloid nucleus (LA) can be considered an interface to sensory systems as shown in Figure 3.7 on the next page. The figure shows that the LA receives projections from a number of neural structures, in particular the septo-hippocampal system and the locus coeruleus. The LA is therefore a site of convergence for neurochemical signals that convey distinctive qualitative information to the amygdala, e.g. novel stimuli are signalled by noradrenaline projections from the LC, and mismatch between stimuli are signalled by glutamate projections from the septo-hippocampal system (from stimulus comparisons occurring in the subiculum).

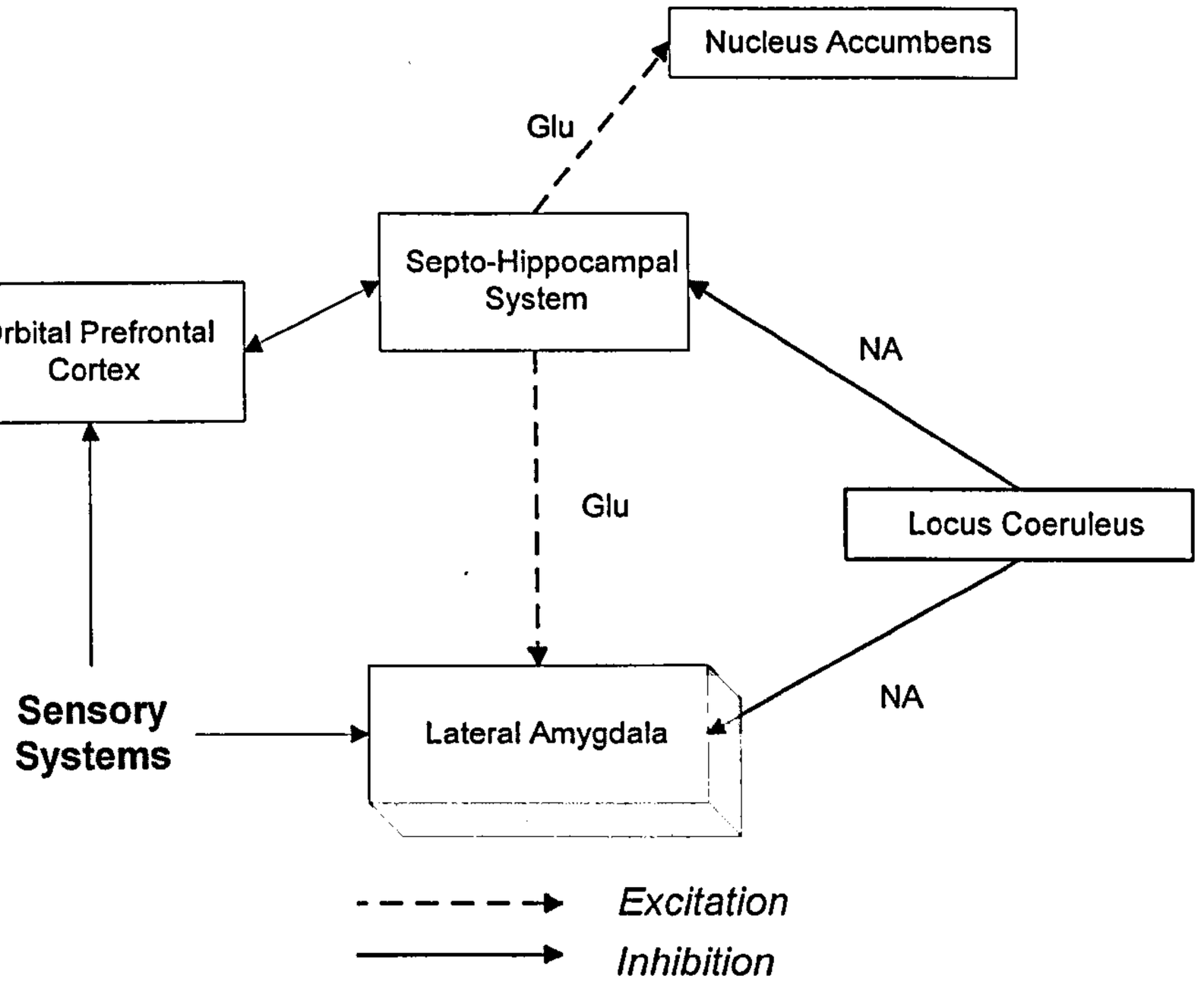


Figure 3.7 - Lateral Amygdaloid Nucleus as Interface to Sensory Systems

The model postulates that the LA is the sensory interface to the rest of the amygdaloid complex, which is consistent with the work of Bordi & LeDoux [1992] who found that neurons in the LA displayed sensory tuning (receptive field) properties as a result of direct association with various sensory areas. The LA receives most its direct sensory inputs from sensory processing areas in the cortex and thalamus, and therefore receives information about stimuli from many sensory modalities. Earlier investigation into auditory processing areas by LeDoux et al. [1990] suggested that the LA has an architecture designed to respond to acoustic events over a specific range of frequencies. Certain neurons in the LA responded selectively to "loud" events, but were insensitive to low levels of auditory stimulation. It is possible that this is a hard-wired property, but more likely that neural plasticity and a learning mechanism is in operation [LeDoux et al., 1990]. The potentiation of neurons in this nucleus could be influenced by increasing the

propensity to fire of neurons associated with threatening (conditioned) stimuli, and decreasing the propensity to fire of neurons associated with non-threatening (unconditioned) stimuli. This requires that the amygdaloid complex is aware of what constitutes a threatening stimulus, i.e. that reward information is received at the same time. The model postulates that rewards are the function of the basolateral amygdaloid nucleus, which is described in the next section.

3.7.2 - The Basolateral Amygdaloid Nucleus and the Reward System

The basolateral amygdaloid nucleus (BLA) can be considered as the neural centre for the reward system in this part of the brain as shown in Figure 3.8 on the next page. The model postulates that the BLA is responsible for forming associations between sensory stimuli that arrive at the lateral nucleus, and reward information that is available to the BLA. This assumes there is a connection between the lateral and basolateral nuclei, and is supported by evidence showing that particularly high concentrations of noradrenaline have been found in the basolateral amygdaloid nucleus [Sarter & Markowitsch, 1985]. The connection between the orbital prefrontal cortex and BLA is significant because Rolls [1990] argues that the orbital prefrontal cortex is involved in rapid adjustments of behavioural responses made to stimuli when their reinforcement value changes, such as in reversal tasks. He describes experiments that show single neurons in the orbital prefrontal cortex respond to non-rewards after reversal because of their previous association with rewards, and suggests that this response reflects the situation where expected rewards no longer correlate with actual rewards. This information is necessary for adapting behavioural responses when reinforcement contingencies are changed, as shown by the fact that monkeys with lesions to the orbital prefrontal cortex are unable to change their responses and show perseverative behaviour [Rolls, 1990]. Similar results have been drawn

from studies with rats [Kesner, 1992]. The model suggests that the connections between the orbital prefrontal cortex and the BLA hold information about actual rewards, and this information would be useful in the tasks described above.

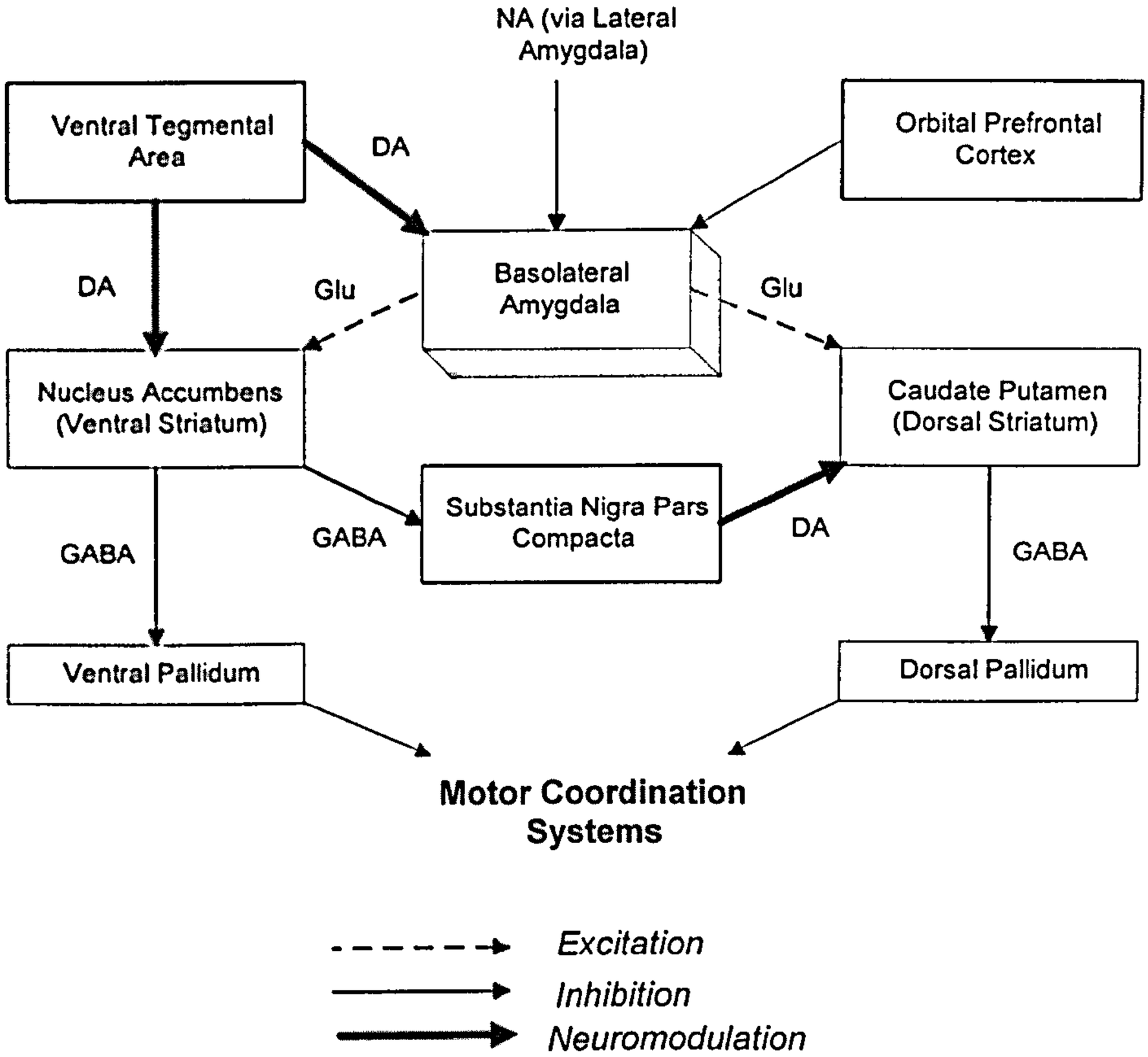


Figure 3.8 - Basolateral Amygdaloid Nucleus and the Reward System

Cools et al. [1991] suggest that noradrenaline has a 'gating' action on information reaching the ventral striatum, and selectively modulates the inputs from the BLA and hippocampus to the nucleus accumbens. Dopamine projections from the ventral tegmental area convey information about rewards directly to the BLA (as well as to the ventral striatum). The BLA therefore knows about actual reinforcements, and is able to compare this to previous reward information available from the orbital prefrontal cortex. The stimuli to which these rewards relate would be provided by the LC through the LA, and knowledge

of these stimuli would be enhanced by the presence of noradrenaline at the BLA when situations requiring vigilance (perhaps the result of a change to the stimulus and reinforcement contingency) are encountered. When relearning is required, this would lead to the BLA exerting an influence on processing in the ventral striatum by releasing glutamate, which increases the release of dopamine at the ventral and dorsal striatum.

The ventral striatum (specifically nucleus accumbens) acts as a 'switch', and can activate motor behaviours using either GABA projections to the thalamus, or by indirect GABA projections to the dorsal striatum via the substantia nigra pars compacta. It has been suggested that this mechanism is provided by means of the 'selective disinhibition' of projections to the dorsal striatum from the nucleus accumbens [Chevalier & Deniau, 1990]. The model therefore suggests that relearning processes can be influenced in the ventral striatum by two different pathways :-

- Noradrenaline signals indicating novelty into the BLA will release glutamate and increase the postsynaptic activity of dopamine neurons in the ventral striatum. This novelty represents a change to the association between stimuli and rewards, and perhaps indicates that there is a mismatch between a stimulus and an expected reward. This signal would also be present the first time a stimulus and reward contingency is detected.

- The subicular glutamatergic projection to the ventral striatum will indicate when sensory mismatch has been detected in the subiculum. This is also a form of novelty, but represents the mismatch between stimuli and their relationship with other stimuli.

Both of these pathways increase the activity of dopamine neurons in the ventral striatum by increasing the activity of glutamate projections to the ventral striatum. This is supported by the work of Burns et al. [1996], who demonstrate increased aversive responses to novelty and greater exploration, and suggest that these are modulated by the interaction of glutamate and dopamine in the nucleus accumbens. This interaction is between dopamine originating from the ventral tegmental area, and glutamate from the BLA, subiculum and prefrontal cortex. Learning in the ventral striatum is therefore amplified by the involvement of the amygdala, and this feature is useful when relearning processes are required. This aspect of the model has implications for the relearning problem because the amygdaloid complex is aware of novel and unexpected features of situations it has encountered, and can relate this to the expectation of reinforcement. Attention can therefore be focused when these situations arise, rather than initiating a strategy of total relearning. The switching mechanism provides an opportunity to use previously learned motor patterns in the dorsal striatum by terminating the current behaviour, and perhaps initiating exploratory behaviours.

3.7.3 - The Central Amygdaloid Nucleus and Motor Activity

The model postulates that at the same time as glutamate is released to influence the activity in the ventral striatum, acetylcholine is released to affect the activity in a number of other systems. This assumes that the BLA has some way of passing this information to the central amygdaloid nucleus (CeA), and glutamatergic projections from the CeA to the mediodorsal thalamus suggest that there are indeed glutamatergic projections between the BLA and CeA [Amaral et al., 1992]. At the same time as relearning processes are initiated, the CeA exerts an influence over perceptual and autonomic systems to inhibit behaviour (e.g. freezing) and increase attention and arousal (e.g. increase heart rate) as described by

Gray [1995]. Memory consolidation can also be influenced by the connections through the thalamus to the cortex, and to the entorhinal cortex. The influence of the central amygdaloid nucleus on motor systems is summarised in Figure 3.9.

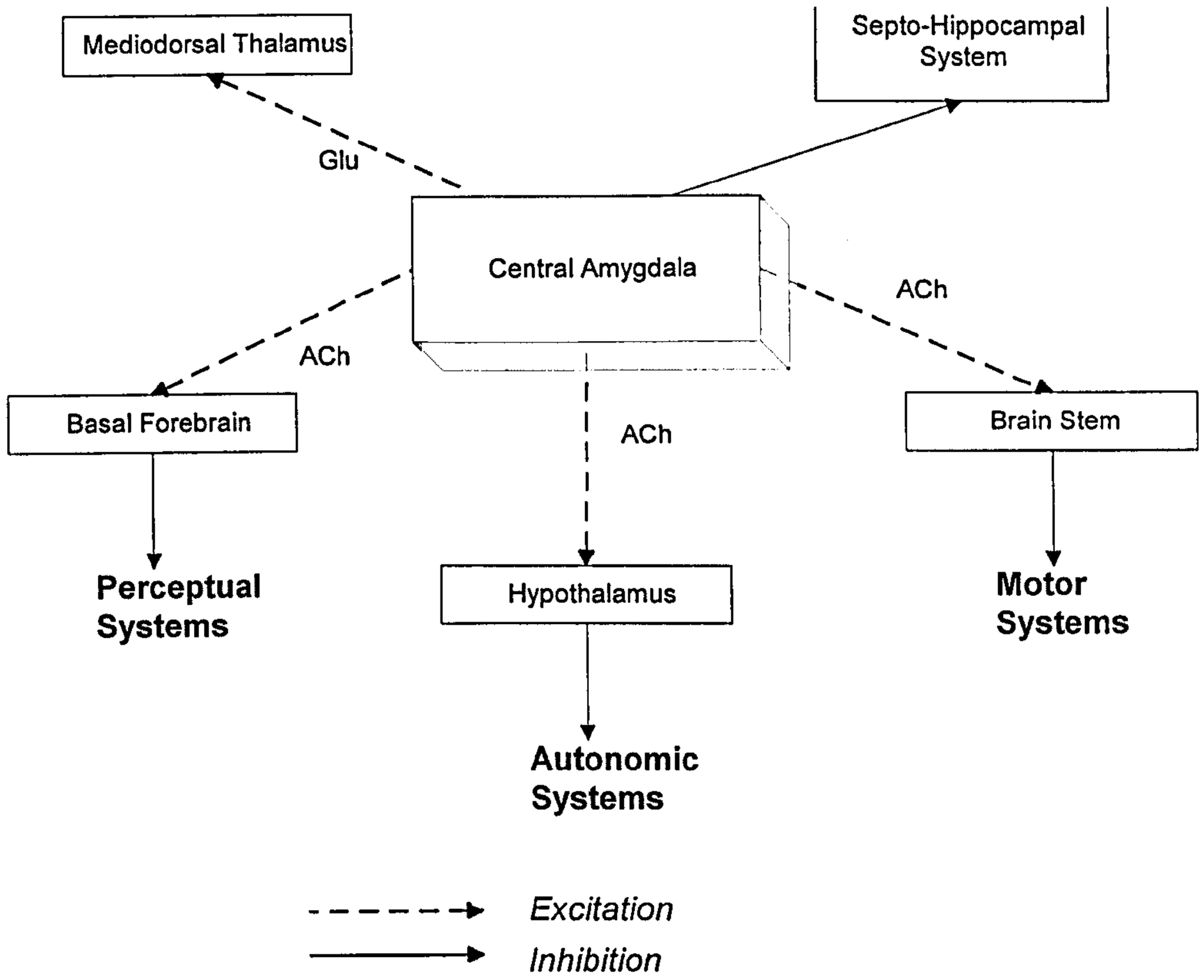


Figure 3.9 - Central Amygdaloid Nucleus and Motor Activity

This aspect of the model is useful because it provides a means by which learning can be enhanced in a number of brain areas when relearning is required. This would mean that an affective bias is added to the information content in those areas, and this is important in the context of intelligent behaviour as described earlier in the chapter.

3.8 - Summary

This chapter looked at reinforcement learning and relearning from a biological perspective, and discussed how the amygdaloid complex is involved. A conceptual model was proposed that describes how the nuclei of the amygdaloid complex receive information about stimuli and rewards, and project this information to structures in the basal ganglia. Other structures (e.g. the septo-hippocampal system) have access to information about the relationships between stimuli, and project to the same areas of the basal ganglia. The fact that two systems have access to the same information is important, and allows either system to influence the learning of motor behaviour in the basal ganglia. In particular, the two systems can detect novelty in terms of a sensory mismatch (the septo-hippocampal system) and in terms of expected rewards (the amygdaloid complex and its connection with the prefrontal cortex, locus coeruleus, ventral tegmental area). A 'switching' mechanism was described that allows the learning of motor skills in the ventral striatum to select between previously learned patterns of motor behaviour held in the dorsal striatum, and may be achieved by selective disinhibition of required connections. The model proposes that this is the role of glutamate projections to the ventral striatum, which then influences the dorsal striatum via the substantia nigra. The model also proposes that the amygdaloid complex can influence a number of other systems when relearning is required, thus providing a means of initiating various attentional mechanisms.

It needs to be emphasised that the conceptual model proposed in this chapter is speculative in nature, and is based on the evidence presented from a number of experimental studies and critical reviews. This is the knowledge currently available, and the processes described are still not fully understood. The role suggested for the neurochemicals dopamine and glutamate is only hypothetical, and any proposals made by the model have to be taken as tentative. As more knowledge is acquired and the

fundamental processes better understood, it is likely that the model will need to be revised and further developed. However, the model is useful because it may provide inspiration for finding mechanisms to address the relearning problem. The next chapter looks at how such inspiration can suggest modifications to the adaptive critic system by simple abstraction of some of the biological mechanisms described in this chapter.

Chapter Four

Biological Inspiration Applied to the Adaptive Critic System

4.1 - Introduction

Chapter Three presented a conceptual model to account for reinforcement learning and relearning processes in the brain. The conceptual model is based on the hypothesis that two neural structures are able to influence learning and relearning in the basal ganglia: the septo-hippocampal system and the amygdaloid complex. According to the model, the septo-hippocampal system exerts its influence on the basis of observed differences between sensory stimuli and can thus detect when relearning is necessary. Similarly, the amygdaloid complex exerts its influence on the basis of observed differences to reward information, and can thus detect when reinforcement contingencies have changed and relearning is required. The amygdaloid complex receives information from the ventral tegmental area about external reinforcement, and from the locus coeruleus indicating novel stimuli.

Figure 4.1 summarises the reinforcement aspects of the conceptual model in terms of the neurochemical substances involved, and indicates in very crude functional terms the interaction between the amygdaloid complex, the orbital prefrontal cortex, the locus coeruleus and structures of the basal ganglia. The amygdaloid complex is shown as a single structure even though in Chapter Three it was argued that its functionality is provided by individual nuclei within the amygdaloid complex. The intrinsic connections of the amygdaloid complex are still under investigation, and so for the biological inspiration purposes of this thesis, only the functional role of the amygdaloid complex will be

considered. It is therefore shown as a single structure. The influence of the septo-hippocampal system is omitted from the diagram because (according to the conceptual model) it is not directly involved in generating and using reinforcement signals which are the basis for reinforcement learning in the adaptive critic system described in Chapter Two.

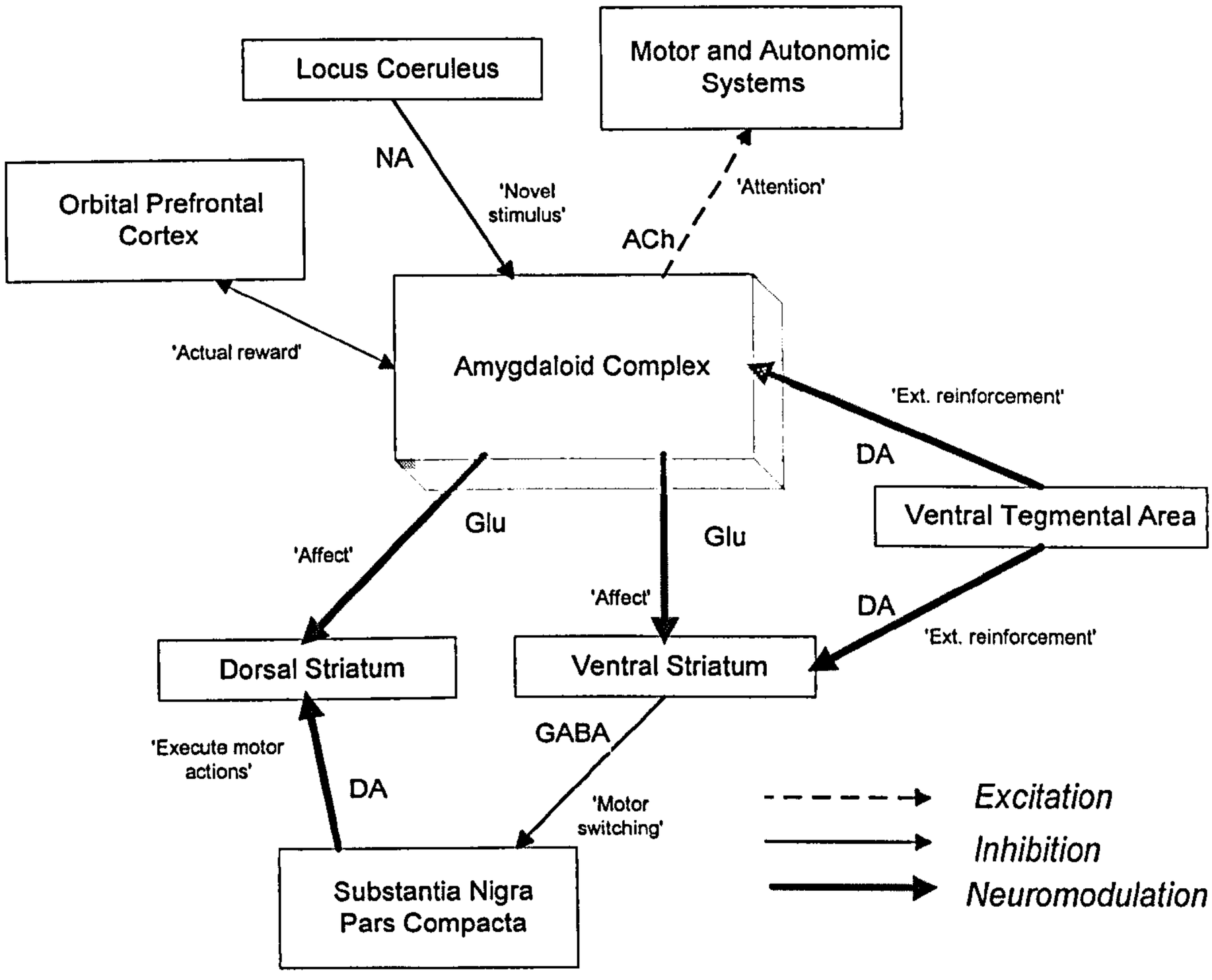


Figure 4.1 - Summary of the Conceptual Model

Figure 4.1 shows that external reinforcement information arrives at the amygdaloid complex by means of dopamine signals originating in the ventral tegmental area. The conceptual model suggests that this external reinforcement information is stored as actual reward by means of reciprocal connections that exist between the amygdaloid complex and the orbital prefrontal cortex. In addition, the locus coeruleus provides the amygdaloid complex with information about novel or significant stimuli by means of noradrenaline

signals. This information may be part of a much more complex novelty detection and exploratory mechanism involving the LC and its connecting structures as suggested in Chapter Three. The significance of the pathway from the LC to the amygdaloid complex is that it provides a means by which the signal/noise ratio can be 'tuned' for sensory stimuli arriving at the amygdaloid complex [Oades, 1985]. Knowledge about novel or unexpected stimuli is therefore available at the amygdaloid complex. The amygdaloid complex is thus able to 'affect' learning in the basal ganglia by increasing the dopamine activity in both the ventral and dorsal striatum, enabled by glutamate projections from the amygdaloid complex to both these structures. As described in Chapter Three, this would lead to increased 'switching' between motor programs in the ventral striatum, and the facilitation of motor action selection in the dorsal striatum. The diagram does not detail all the projections to the basal ganglia that enable this to be achieved, but it does detail the projections from the amygdaloid complex that are of interest. These projections are marked as neuromodulatory, although glutamate is known to be excitatory [Hestenes, 1992]. The reason for this is that the effect on basal ganglia structures could be to either increase or decrease their activity depending on their activity at that particular time. In situations requiring relearning, this effect would be to enhance the learning of particular motor actions (dorsal striatum) and increase switching between sequences of these actions (ventral striatum). At the same time, attention mechanisms would be triggered in structures efferent to the amygdaloid complex by means of acetylcholine projections as described in Chapter Three. This activity would be triggered by the amygdaloid complex becoming aware of a change to the reinforcement schedule, and this could be as a result of :-

- a novel stimulus receives an external reinforcement

- a stimulus that previously received a reinforcement no longer receives that reinforcement
- a stimulus that previously did not receive a reinforcement now receives a reinforcement.

It can thus be seen that the apparent role of the amygdaloid complex (according to the conceptual model) is to enhance the activity in the basal ganglia when the above situations occur. This would help to solve part of the relearning problem because situations requiring relearning on the basis of a change to the reinforcement schedule can be detected and acted upon, and this may provide inspiration for developing intelligent control mechanisms. This thesis has concentrated on the adaptive critic system, and the purpose of this chapter is to describe how this inspiration can be applied to the adaptive critic system to enable it to better address the relearning problem. This requires that the adaptive critic system is related to the neurophysiological mechanisms on which many researchers (e.g. Werbos [1995], Barto [1995]) argue that it has been derived. This was the essence of the work conducted by Houk et al. [1995b] which will now be described.

4.2 - The Houk et al. Model

Houk et al. [1995b] developed a model to explain how structures in the basal ganglia are able to generate and use signals that predict reinforcement by means of 'striosome' compartments in the striatum, and the signalling properties of dopamine (DA) neurons. Houk et al. attempted to relate the architecture of their model to the original adaptive critic system that Barto et al. [1983] used to solve reinforcement learning problems. There appears to be a correlation between the discharge properties of DA neurons in the Houk et al. model, and the effective reinforcement signal generated by the adaptive critic when learning with terminal primary reinforcement [Houk et al., 1995b].

Barto [1995] discussed the similarity between the adaptive critic system and the basal ganglia in detail, describing the adaptive critic as "*a device that learns to anticipate reinforcing events*". A correspondence has therefore been established between the adaptive critic system and basal ganglia models of reinforcement learning.

The Houk at al [1995b] model proposes that dopamine (DA) neurons in the basal ganglia acquire the ability to predict reinforcement, and describes how outputs from these neurons are used to reinforce behaviours that lead to primary reinforcement. The DA neurons have reciprocal connections with spiny neurons in the striatum, and these form modules called 'striosomes'. The same DA neurons are also connected to other spiny neurons in the striatum, and these form modules called 'matrisomes'. Houk et al. proposed that striosome modules use DA inputs to learn how to detect contexts that precede reinforcement by a short time interval, and these acquired responses are then used to control their own DA input. Through this recursive mechanism, DA neurons are able to learn to detect earlier and earlier predictors of reinforcement. The same DA signals also reinforce the spiny neurons in the matrisomes such that they learn to detect and register regular contexts useful in the planning and control of motor behaviour.

The following sections describe the organisation of striosome and matrisome modules, and relate this to the adaptive critic system. Striosome modules are able to recursively generate signals that predict future reinforcement, and this will be explained in terms of how these signals correspond to the effective reinforcement signal (or TD error from Chapter Two) produced by the 'critic' in the adaptive critic system.

4.2.1 - Organisation of Striosomes and Matrisomes

The striatum is made up of medium-sized spiny neurons, each receiving cortical input and sending this outward to the pallidum and substantia nigra [Goldman-Rakic &

Selemon, 1990]. The striosomes and matrisomes are composed of these spiny neurons, but the targets to which these neurons project is very different [Graybiel & Kimura, 1995]. Spiny neurons in striosomes project to DA neurons in the substantia nigra and ventral tegmental area, whereas spiny neurons in matrisomes only project to output neurons in the globus pallidus and substantia nigra pars reticulata. The organisation of striosome and matrisome modules is illustrated in Figure 4.2, which shows how each of these modules may be connected with neurons in the cortex, various thalamic nuclei, and other basal ganglia structures.

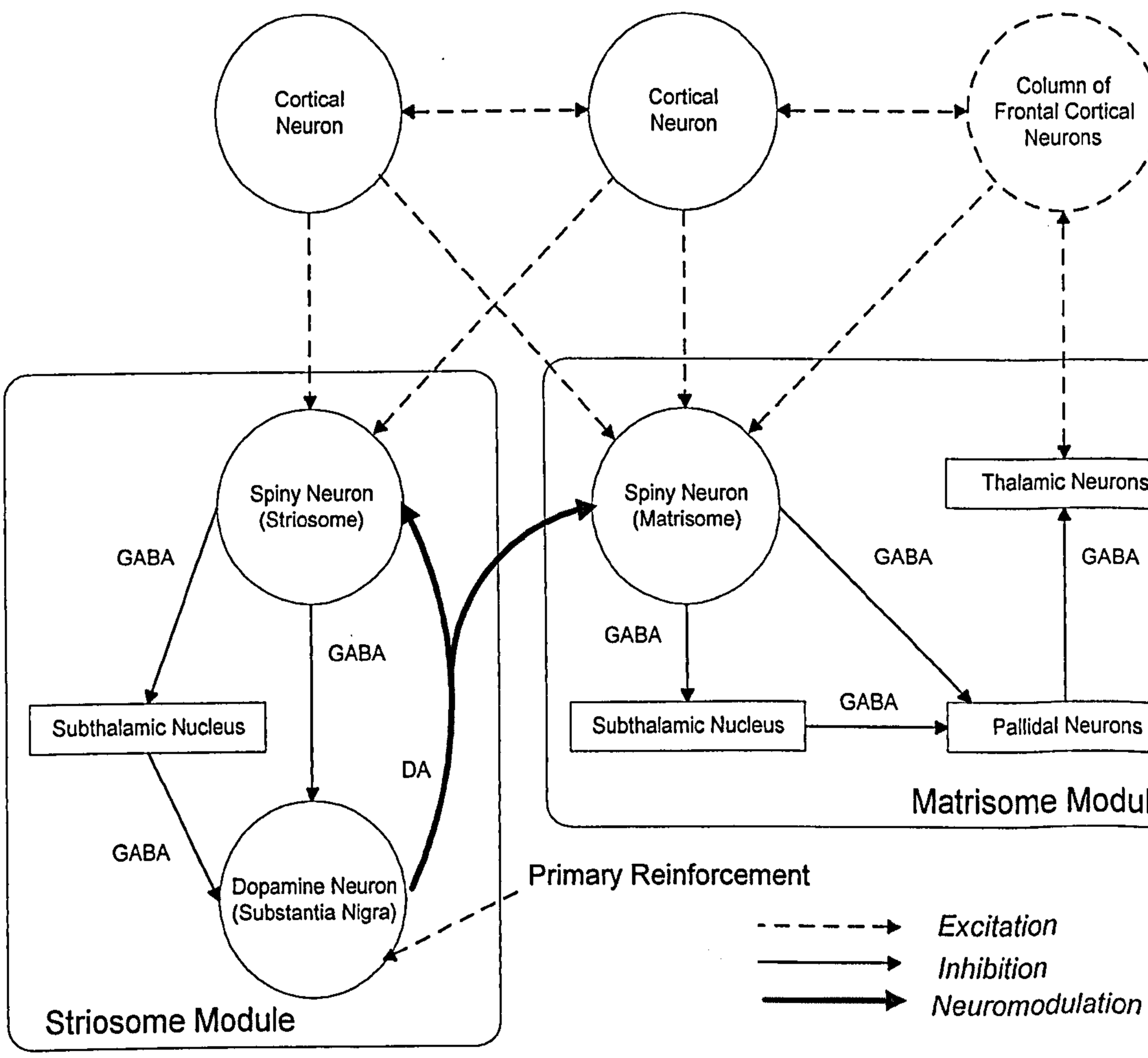


Figure 4.2 - Modular Organisation (Adapted from [Houk et al., 1995b])

The diagram shows two out of thousands of neurons located in different areas of the cortex that function as context detectors. These neurons send convergent inputs to spiny neurons in both striosome and matrisome modules, which means that spiny neurons in both modules receive organised, convergent input from widespread areas of the cerebral cortex [Graybiel & Kimura, 1995]. The Houk et al. [1995b] model suggests that this makes the striatum ideally suited for the recognition of complex patterns stored by cortical neurons. According to Houk et al., the DA neuron in the striosome module (the DA neuron itself is actually in the substantia nigra) receives three input projections :-

- The first projection is from the spiny neuron in the striosome. This projection is GABAergic and thus inhibitory.
- The second projection is also from the spiny neuron in the striosome, but projects indirectly through the subthalamic nucleus to the DA neuron. This projection has a net disinhibitory action as the subthalamic nucleus projections are also GABAergic [Chevalier & Deniau, 1990].
- The third projection is labelled 'Primary Reinforcement'. Houk et al. suggest that this is likely to be an excitatory projection from the lateral hypothalamus which relates to primary reinforcement of an appetitive nature.

Figure 4.2 also shows that the DA neuron in the striosome module projects back to the same spiny neuron that sends it input, and Houk et al. [1995b] suggest that this enables the DA neuron to make progressively earlier predictions of reinforcement. There is substantial evidence to suggest that DA neurons located in the ventral tegmental area and the substantia nigra pars compacta play an essential role in the primary reinforcement of behaviour, and in guiding preparatory behaviour on the basis of the likelihood of receiving

subsequent reinforcements [Houk et al., 1995b]. The work of Schultz et al. [1995] provides such evidence by demonstrating that DA neurons selectively respond to only a limited range of stimuli, the major stimuli being reward-related stimuli that indicate primary reinforcement. Schultz et al. report that novel or unexpected stimuli can also trigger DA neuron responses, perhaps because these stimuli are potential rewards or reward predictors that can be included as a class of reward-related events. The DA neurons eventually lose their ability to respond to reward stimuli once the stimuli have become valid predictors of reward. A precise neurophysiological explanation for why these responses disappear is still unclear, but Schultz et al. [1995] suggest that this is because the responsibility for signalling reward is taken over by neurons located elsewhere, possibly in the amygdala. This is consistent with the conceptual model in Chapter Three because actual rewards are recorded by the amygdala, and such information is only of significance in the basal ganglia when something changes and relearning is required. The thesis has argued that this is the responsibility of the amygdaloid complex.

4.2.2 - Relating the Model to Temporal Difference Methods

Thus far the Houk et al. [1995b] model has established that the adaptive critic and basal ganglia are similar in terms of structure. The underlying functionality of basal ganglia mechanisms needs to be related to the computational operations performed by the adaptive critic system. Houk et al. have described how the DA neuron in the striosome module learns to fire in response to one context predictive of reinforcement, and then uses the pathway back to the spiny neuron to reinforce itself for firing to an even earlier context that predicts reinforcement. This mechanism is essentially the same as learning to predict reinforcement described as Temporal Difference Methods in Chapter Two. A detailed neurochemical explanation of this mechanism is beyond the scope of this work, and the

reader is referred to Houk et al. [1995b] for more detail. The important point is that Houk et al. refer to a neurochemical mechanism that enables DA neurons to respond to predictions of reinforcement without having to rely on primary reinforcement signals. The striosome module can thus function as a secondary reinforcer, and is analogous to the ability of the adaptive critic to predict reinforcement using successive predictions. The basic idea is to let the predictions of reinforcement generated by the 'critic' serve as surrogate (or secondary) reinforcers for controlling the 'actor', which generates the control actions. Houk et al. [1995b] have thus suggested that the critic is functionally equivalent to the striosome module, and that the actor is functionally equivalent to the matrisome module. The critic provides the actor with an effective reinforcement signal ($\hat{r}_t$) which influences how the actor learns control actions. This is shown in Equation (7) repeated from Chapter Two, where P_t is the prediction at time t, and P_{t-1} is the prediction at the previous time step. These predictions are defined in equations (5) and (6) of Chapter Two.

$$\hat{r}_t = r(x_t) + \gamma \; P_t - P_{t-1} \tag{7}$$

The Houk et al model [1995b] suggests that $\hat{r}_t$ is carried by the DA projection from the DA neuron in the striosome module to the spiny neuron in the matrisome module. This dopaminergic projection 'neuromodulates' spiny neurons in the matrisome. The Houk et al. model also suggests that $r(x_t)$ corresponds to the primary reinforcement signal of an appetitive nature from the lateral hypothalamus. The predictions P_t and P_{t-1} are generated by the spiny neuron in the striosome module and carried by GABAergic projections to the DA neuron in the striosome module. The correspondence between the Houk et al. [1995b] model (using striosomes and matrisomes) and the critic and actor of the adaptive critic

system has thus been related to Equation (7). The learning coefficient 'γ' has not yet been discussed, and will be considered later in this chapter.

4.3 - An Alternative Biological Basis for the Adaptive Critic System

It is at this point that the conceptual model described in Chapter Three can be considered. There is a similarity between the Houk et al. [1995b] model and the adaptive critic system as described in the previous section. There are also similarities between the conceptual model described in Chapter Three and the Houk et al. model [1995b], which can be summarised as follows :-

- Both the Houk et al. [1995b] model and the conceptual model suggest the striatum is the place where motor actions are related to reinforcements
- The predictions generated by the basal ganglia are used for learning sequences of motor actions, and both models suggest that this learning occurs in the striatum
- Both models implicate dopamine as a neuromodulator which affects the learning of motor action sequences.

There are, however, some important differences between the two models :-

- The Houk et al model suggests that matrisome modules are used for motor learning, whereas the conceptual model suggests that motor learning occurs in the dorsal striatum
- The Houk et al. model suggests that striosome modules are responsible for generating and storing predictions of reinforcement, whereas the conceptual model suggests that the ventral striatum enables 'switching' between motor action sequences

- The origin of the primary reinforcement signal in the Houk et al. model is the lateral hypothalamus, whereas the conceptual model suggests this signal originates in the ventral tegmental area
- The Houk et al. model only considers reinforcement learning, whereas the conceptual model also considers relearning
- The involvement of the amygdaloid complex is central to the conceptual model, but is not considered at all by the Houk et al. model.

It is possible to use the basal ganglia aspects of the conceptual model to provide an alternative biological basis for the adaptive critic system. This may be achieved by relocating the components of the adaptive critic in the ventral and dorsal striatum such that the ventral striatum may be considered as the critic, and the dorsal striatum may be considered as the actor. This does not detract from the functionality of the Houk et al. [1995b] model, but simply locates this functionality in a different part of the basal ganglia. The ventral and dorsal striatum may well be neurophysiologically different to the matrisomes and striosomes described by Houk et al., and this would require further investigation, but this does not diminish the argument presented in this thesis. The conceptual model argues that the interaction between neurochemical systems is the basis for selective attentional mechanisms influencing a number of brain areas when relearning is required. This interaction is enabled by glutamatergic projections from the amygdaloid complex to both the ventral and dorsal striatum, which was not considered by Houk et al. [1995b] in their model. The conceptual model suggests that learning in the ventral striatum (critic) and dorsal striatum (actor) is affected by glutamate projections from the amygdaloid complex, which leads to increased dopamine activity in these structures. These effects should also be reflected in the adaptive critic system. This is the justification for

introducing the amygdaloid complex as an additional 'amygdala component' in a modified version of the adaptive critic system, as shown in Figure 4.3. The figure shows that the postulated effects of the neurochemical projections from the 'amygdala component' identified by the conceptual model are equivalent to neuromodulatory affects in the modified adaptive critic system.

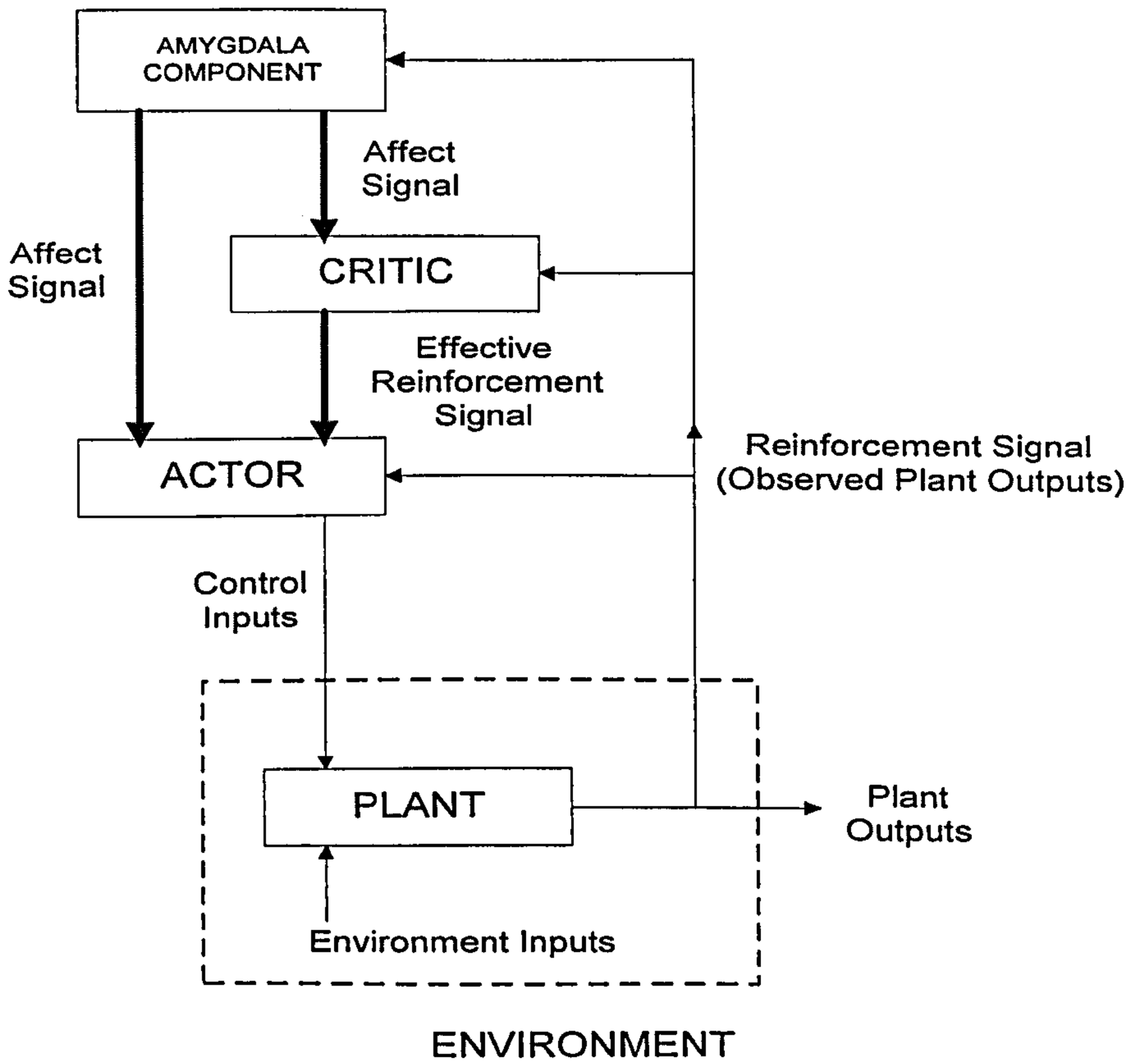


Figure 4.3 - Modified Adaptive Critic System

When novel or unexpected situations are detected (e.g. reinforcing stimuli that have not been encountered before, or reinforcements that do not match predictions), the amygdala component provides a signal to both the actor and critic that will 'affect' the learning of appropriate behaviour. In the adaptive critic system, behaviour is expressed as the control actions determined by the actor. These control actions are influenced by the

effective reinforcement signal which is determined by the critic. Both actor and critic are represented by a system of equations. If the amygdala component affects the learning processes in the actor and critic when relearning is required, then it is argued that this influence will be reflected in the equations that make up the adaptive critic system. These equations represent the computations of the actor and critic, and include a number of learning coefficients (such as 'γ' in Equation (7) above) that influence different aspects of learning. It is suggested that if the amygdala component has a 'neuromodulatory' affect, then this will be manifested by an influence on the learning coefficients. Other researchers have used the modulation of learning coefficients as a method for providing exploration mechanisms, e.g. the gain parameter used for selective attention [Thrun, 1992] described in Chapter Two. Cohen & Servan-Schreiber [1992] described how a simple gain parameter can be used to simulate different neuromodulatory effects at both the biological and behavioural level. They related these effects to the neuromodulatory effects of dopamine in schizophrenia, and argued that these effects could be reproduced by simply changing the gain equally for all units influenced by the neuromodulator. Their focus was on dopamine in the prefrontal cortex, and they showed that increasing the gain in their simulations produced the same cognitive deficits associated with increased dopamine activity in the prefrontal cortex, and decreasing the gain produced the same deficits associated with a decrease in dopamine activity in the prefrontal cortex. The computational modelling of neuromodulation has received much recent interest, and various proposals have been made (e.g. Bower [1993], Myers et al. [1996], Rumelhart [1997], and various papers presented at a special session on neuromodulation held at NIPS'96).

In neurophysiological terms, the effect of neuromodulation can be simplified to an increase (or decrease) in the learning activity at all areas affected by the neuromodulator substance or substances. This obviously depends on whether the effect of these

neurochemical substances is predominantly excitatory or inhibitory. Equation (10) suggests how a learning coefficient can be modulated:

$$coefficient_{t+1} = \kappa * coefficient_t + \omega \qquad (10)$$

In Equation (10), the value of the learning coefficient at time $t + 1$ will be the value of that coefficient at time t plus an increment (ω) that represents an increase or decrease in learning activity. If the neurochemical effect is predominantly excitatory, then the increment will be positive and the value of the coefficient will increase when the amygdala component detects that reinforcement contingencies have changed . If the neurochemical effect is predominantly inhibitory, then the increment will be negative (the same as a decrement) and the value of the learning coefficient will decrease when the amygdala component detects that reinforcement contingencies have changed. It must also be remembered that the effects on learning activity caused by the neuromodulator will decay with time if not maintained, and hence a decay term K has been included. It is assumed that every learning coefficient will have a minimum and maximum activity, and this may be represented as minimum and maximum threshold values for that coefficient. For example, if the neurotransmitter is excitatory (such as with glutamate and acetylcholine), then the increment will be positive and the coefficient will be expected to increase when modulated and decay with time. If the neurotransmitter is inhibitory (such as with GABA), then the increment will be negative and the coefficient will be expected to decrease when modulated and return to its original activity threshold value with time. The following section looks at how neuromodulation might be achieved in the adaptive critic system.

4.4 - Modulation of Learning Coefficients

There are a number of learning coefficients in the equations that make up the adaptive critic system that are candidates for modulation by the amygdala component in the modified adaptive critic system, and this is because the amygdala component influences both the actor and the critic. If the learning coefficients are to be modulated, the question is then *how* and *when* should they be modulated so as to appropriately influence learning? These coefficients would normally be constants fixed before learning. A few studies have investigated how these coefficients affect learning performance, such as Geva & Sitte [1992] who looked at the effect of optimising coefficients in the adaptive critic system for the pole balancing problem. They tried different combinations of fixed constants to see if they could characterise learning performance, the aim being to optimise these constants and achieve the fastest possible learning times. They described this endeavour as a laborious process, and could only conclude from their data that the adaptive critic system is not very sensitive to coefficient setting. This would therefore seem to suggest that modulation of learning coefficients will have little effect. This needs to be questioned, because intuitively, certain coefficients would be expected to have an effect else why would they be included? For example, the value of the 'γ' discount factor affects how successive predictions of reinforcement are discounted by the critic, and hence influences the effective reinforcement signal given in Equation (7):

$$\hat{r}_t = r(x_t) + \gamma \; P_t - P_{t-1} \tag{7}$$

Barto [1995] describes the discount factor as determining how strongly predictions of future primary reinforcement should influence current actions. Any amount of primary

reinforcement that is delayed by one time step is worth a fraction (γ) of that same amount of undelayed primary reinforcement. When $\gamma = 0$, the current prediction of future reinforcement P_t is not taken into account, and learning depends more on the actual reinforcement $r(x_t)$ and the short-term consequence of an action. This may be considered a 'tactical' learning objective. As 'γ' increases towards 1, the current prediction of future reinforcement becomes more significant because the delay is taken into account. This may be considered a 'strategic' learning objective because actions are reinforced by both their long-term and short-term consequences. In the relearning problem (where the reinforcement contingency has changed), the delayed consequence of an action should be less important because this may no longer be applicable. It is argued that the learning objective becomes more tactical and less strategic when relearning is required. This means that the value of the discount factor should tend towards zero when relearning is required. It is therefore suggested that modulating the 'γ' learning coefficient in this way may prove beneficial when addressing the relearning problem.

Another example is the 'λ' coefficient, which is used in TD(λ) as given by Equation (9) repeated from Chapter Two:

$$\bar{a}(x_{t+1}) = \lambda \bar{a}(x_t) + (1-\lambda) \tag{9}$$

The value of the 'λ' coefficient was investigated by Sutton & Barto [1995] whose work addressed the question of whether reinforcement learning methods are better when they learn on the basis of actual outcomes (i.e. $\lambda = 1$) or on the basis of interim estimates (i.e. when $\lambda < 1$). They argued that the former are better when function approximators are used, but the latter are thought to achieve better learning rates. This question had not been put to an empirical test using function approximators, and so Sutton & Barto [1995]

presented preliminary results from such a test on a variety of different problems including the pole balancing problem. They plotted 'λ' against performance for each test, i.e. 'λ' versus 'Failures per 100,000 time steps' for the pole balancing problem. Their results show that for the pole balancing problem (and on other problems not reported here) the performance was an inverted U-shaped function of 'λ', with performance rapidly degrading as 'λ' approaches 1 where the worst performance is obtained. These results are based on the 'λ' coefficient being fixed before a test, and remaining constant (i.e. not modulated) for the duration of that test. This is illustrated by Figure 4.4, which qualitatively shows the unpublished data described by Sutton & Barto [1995]:

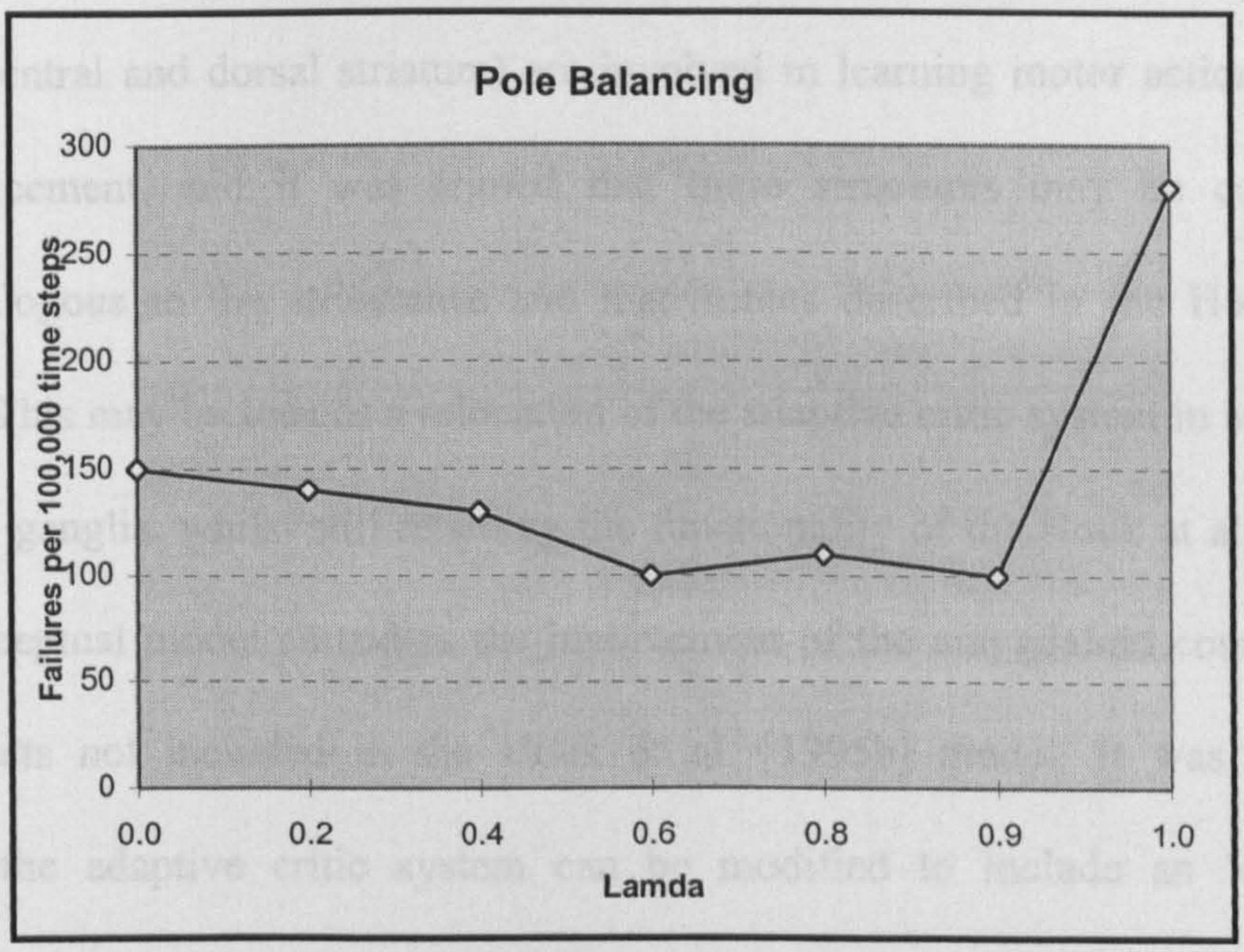


Figure 4.4 - 'λ' in Pole Balancing (Adapted from Sutton & Barto [1995])

There are no results for *modulating* the 'λ' coefficient during learning. It is therefore suggested that modulating 'λ' may prove beneficial for addressing the relearning problem because this coefficient does appear to have an important influence on

performance. The next chapter will consider how this can be achieved experimentally in the context of relearning by modulating 'λ' and 'γ', as well as other coefficients.

4.5 - Summary

This chapter described the Houk et al. [1995b] model which suggests that striosome and matrisome modules in the basal ganglia are able to generate signals that predict reinforcement. The similarity between the Houk et al. [1995b] model and the adaptive critic system presented in Chapter Two was discussed, showing that. striosome and matrisome modules appear to perform the same functions as the critic and actor respectively. The conceptual model presented in Chapter Three suggests that basal ganglia structures (the ventral and dorsal striatum) are involved in learning motor actions on the basis of reinforcement, and it was argued that these structures may be considered functionally analogous to the striosomes and matrisomes described in the Houk et al. [1995b] model. This may be seen as a relocation of the adaptive critic system in a different part of the basal ganglia, whilst still retaining the functionality of the Houk at al. [1995b] model. The conceptual model considers the involvement of the amygdaloid complex and relearning, aspects not included in the Houk et al. [1995b] model. It was therefore suggested that the adaptive critic system can be modified to include an 'amygdala component' in accordance with the conceptual model, thereby capturing the hypothesised involvement of the amygdaloid complex in relearning. The additional amygdala component records actual reinforcements, and is thus able to detect when reinforcement contingencies have changed, The amygdaloid component thus exerts an influence over both the actor and critic when relearning is necessary, achieved by modulating the learning coefficients in the equations that make up the adaptive critic system. The biological

evidence that supports the conceptual model suggests that modulation by the amygdala component occurs when specific situations are detected, i.e. when novel situations are encountered, or when changes to the reinforcement contingency are detected. This alters the value of the learning coefficients in accordance with the information provided by the amygdala component, i.e. where these situations were detected. The next chapter uses experimental simulation to investigate this, and looks at the effects of modulated learning coefficients on relearning when applied to the pole balancing problem.

Chapter Five

Modulated Coefficients and Relearning: A Pole Balancing Example

5.1 - Introduction

The previous chapter suggested that learning coefficients in the equations that make up the adaptive critic system can be modulated to have an influence that is functionally equivalent to the hypothesised influence of certain brain regions. This may lead to an improvement in the performance of the adaptive critic system on the relearning problem, i.e. the problem of detecting that reinforcement schedules have changed, and then using this knowledge to beneficially influence learning processes. This chapter provides an experimental framework within which the modified adaptive critic system can be tested on an established reinforcement learning control problem. The hypothesis is that the learning coefficients in the adaptive critic equations may be modulated in accordance with the memory of previous reinforcements, which is likely to be of benefit to the performance of the system when reinforcement contingencies change and relearning is required.

It is widely accepted that the pole balancing (or inverted pendulum) problem is a difficult non-linear control task [Geva & Sitte, 1992]. This problem has been extensively studied as an easily understood physical system that readily lends itself to the testing of control algorithms at many different levels of complexity [Larcombe, 1996]. The pole balancing problem can be defined in terms of specified parameters and system components given by Geva & Sitte [1993] that provides an experimental benchmark for computer simulation of the problem. The benchmark does not include the relearning problem, and

therefore this chapter will suggest how the benchmark may be extended to accommodate the relearning problem, i.e. how to change reinforcement contingencies to provide an appropriate problem for the modified adaptive critic system. Barto et al. [1983] showed that the pole balancing problem can be addressed using two neuron-like elements called the Adaptive Search Element (ASE) and the Adaptive Critic Element (ACE). This chapter provides a detailed description of the ASE/ACE implementation, which is the original level one adaptive critic system referred to in Chapter Two. The performance of the ASE/ACE system on the benchmark relearning problem is shown using experimental simulation results. Similar experiments were conducted to investigate the effect of modulating learning coefficients with a modified adaptive critic system on the benchmark relearning problem. Each simulation experiment is described, and the results presented and analysed.

5.2 - Definition of the Problem

The pole balancing problem involves balancing a pole attached vertically to a moveable cart placed on a finite length of track, and this will be referred to as the cart-pole system. Control actions are achieved by applying a one-dimensional force of constant magnitude to the base of the cart, the direction of the force being influenced by the controller. This is known as *bang-bang control*. The general problem is to discover a sequence of binary (left or right) control forces that can keep the system balanced for long periods of time. The only external information available to the control system is a negative reward signal given when the system fails, and this is consistent with the reinforcement learning scheme described in Chapter Two.

The state of the cart-pole system at a given time depends on four variables: the position of the cart (x), the linear velocity of the cart ($\dot{x}$), the angle of inclination of the

pole (θ), and the angular velocity of the pole ($\dot{\theta}$). A representation of the cart-pole system showing x, θ, and the control force F is given in Figure 5.1.

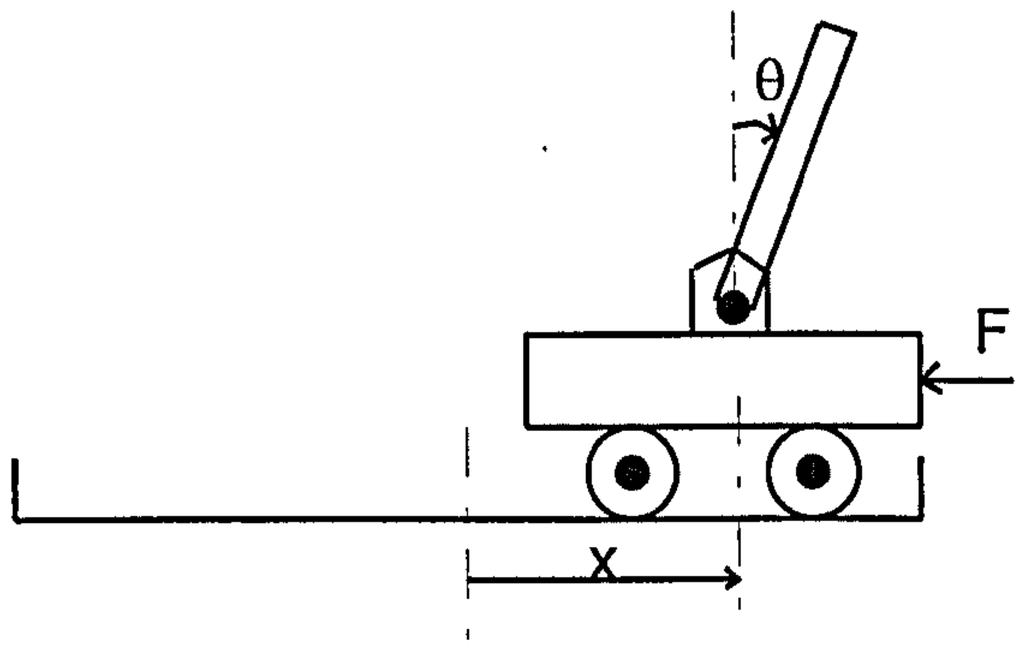


Figure 5.1 - The Cart-Pole System (From Bapi et al. [1997])

5.2.1 - Dynamics of the System

The dynamics of the system are calculated with equations based on the following parameters: the mass of the pole (m_p), the mass of the cart (m_c), the length of the pole (l), the control force (F), and the acceleration due to gravity (g). These equations are given in Figure 5.2.

$$\frac{d^2\theta}{dt^2} = \frac{g\sin\theta - a\cos\theta - \mu_p \dot{\theta}^2 l \cos\theta \sin\theta}{l\left(\frac{4}{3} - \mu_p \cos^2\theta\right)}; \text{ where } a = \frac{F}{m_p + m_c} \text{ and } \mu_p = \frac{m_p}{m_p + m_c}$$

$$\frac{d^2x}{dt^2} = \frac{\frac{4}{3}a + \left(\frac{4}{3}\dot{\theta}^2 l - g\cos\theta\right)\mu_p \sin\theta}{\left(\frac{4}{3} - \mu_p \cos^2\theta\right)}$$

Figure 5.2 - Dynamical Equations of the System (From Bapi et al. [1997])

The equations are integrated numerically using Euler's method with a time step of 0.02 s, and are the revised equations used by Bapi et al. [1997] to omit friction as specified by Geva & Sitte [1993]. The omission of friction is discussed in Section 5.4.

5.2.2 - The Reward Signal

The pole balancing system is said to have failed if pre-specified limits for either the *failure angle* (the angle of the pendulum from the upward vertical about the centre of the track) or the *failure length* (the position of the cart on the track) is exceeded. When this situation occurs, the learning control system receives a negative reward signal set to -1 which is used for learning. The parameters of the controller are only updated when this signal is received, which means that the controller must evaluate its intermediate actions in the absence of continuous reinforcement.

5.3 - The ASE/ACE System

The ASE/ACE system was originally developed by Barto et al. [1983] to show how single neurons can be used to solve complex learning control problems. This system is the same as the level one adaptive critic where the ASE is the actor, and the ACE is the critic. The Barto et al. implementation combines a number of different techniques to achieve the goal of learning to balance the pole using reinforcement learning, and these techniques include state space quantisation and temporal difference prediction.

5.3.1 - State Space Quantisation

The state space quantisation scheme used in the ASE/ACE system is designed to divide the problem space into 162 regions (states) based on the quantisation of the four

system variables: the position of the cart (x), the linear velocity of the cart ($\dot{x}$), the angle of inclination of the pole (θ), and the angular velocity of the pole ($\dot{\theta}$). The regions produced by this state quantisation scheme are shown in Table 5.1. Square brackets indicate that the value is included in a region, parentheses indicate that the value is not included in that region

Variable	Range	Region
	[-2.4, -0.8)	1
x (m)	[-0.8, 0.8]	2
	(0.8, 2.4]	3
	(-∞, -0.5)	1
$\dot{x}$ (m/s)	[-0.5, 0.5]	2
	(0.5, +∞)	3
	[-1.57, -0.21)	1
	[-0.21, -0.02)	2
θ (rad)	[-0.02, 0.00)	3
	[0.00, 0.02]	4
	(0.02, 0.21]	5
	(0.21, 1.57]	6
	(-∞, -0.87)	1
$\dot{\theta}$ (rad/s)	[-0.87, 0.87]	2
	(0.87, +∞)	3

Table 5.1 - State Space Quantisation Scheme (From Bapi et al. [1997])

The ASE/ACE system uses a decoder to convert the input state vector (made up of the regions corresponding to each of the four system variables) into a 162-bit binary number (d) that denotes which state the system is currently in. This binary number has a unit value for the state that the input vector belongs to, and a zero for the remaining bits. Each of the decoder states is connected to the ASE and the ACE by a set of weights, u_i and v_i, respectively. In other words, each of the 162 system states has its own ASE and ACE weights, u_i and v_i respectively. These weights are set to zero initially. The configuration of the ASE/ACE system is shown in Figure 5.3.

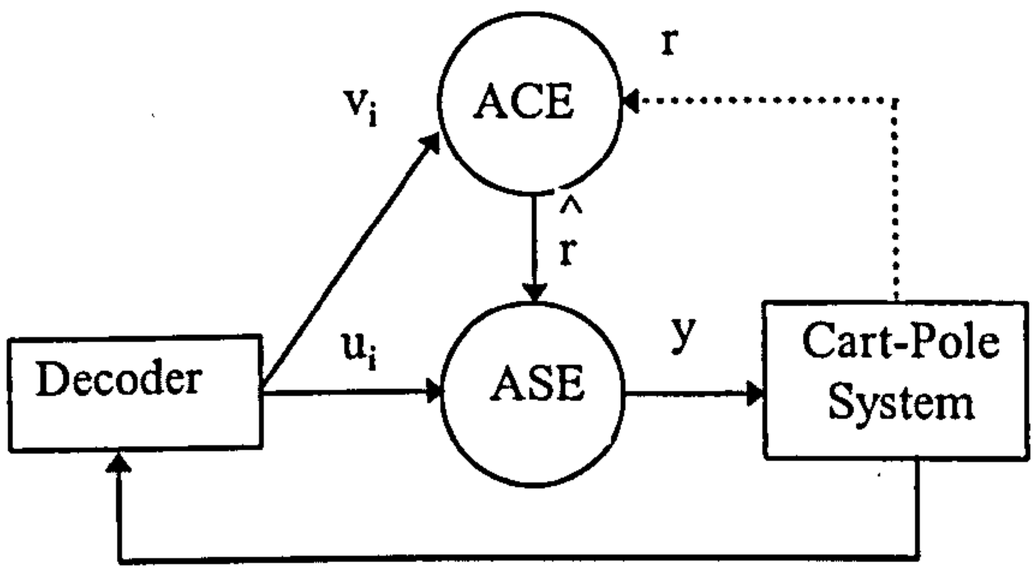


Figure 5.3 - The ASE/ACE System (From Bapi et al., [1997])

5.3.2 - The Associative Search Element (ASE)

The ASE is responsible for generating a binary control action (y) that designates the direction of a control force (F) to be applied to the left or right of the base of the cart. The force is of constant magnitude (10.0 N), and influences the next state of the system in accordance with the equations of motion shown in Figure 5.2. Equation (11) shows how the control force (F) at time t is calculated:

$$F = y(t) * 10 \tag{11}$$

The ASE control action (y) is calculated according to Equation (12):

$$y(t) = g[\sum_i u_i(t)d_i(t) + noise(t)]; \tag{12}$$

$g[w] = +1$, when $w \geq 0$ (right control force)

$g[w] = -1$, when $w < 0$ (left control force)

The control action (y) is simply an indication of the direction of the force, and depends on the weight $u_i(t)$, and the state of the system at time t, given by $d_i(t)$. Initial control actions from the ASE are random because of the noise term (*noise(t)*), which is a Gaussian noise term with zero mean and standard deviation of 0.01. This small term enables the controller to explore the state space in the absence of a known control action. Learning results from updating the ASE weights (u_i) so that eventually the weight is able to overcome the random noise term and is therefore able to generate a known (non-random) control action. Equation (13) shows how the ASE weights are updated:

$$u_i(t+1) = u_i(t) + \alpha \hat{r}(t) e_i(t) \tag{13}$$

In Equation (13), 'α' is a positive learning constant, and $\hat{r}(t)$ represents the internal reinforcement signal provided to the ASE. The term $e_i(t)$ represents the ASE eligibility trace which keeps track of time elapsed since the last visit to a particular state, and every state has an ASE trace associated with it. This trace mechanism therefore helps to solve the temporal credit assignment problem because any state that has not been visited for a long time will have a low $e_i(t)$ value, and its contribution will be almost negligible. The calculation of the ASE trace is given in Equation (14):

$$e_i(t+1)=\delta e_i(t)+(1-\delta)y(t)d_i(t) \tag{14}$$

The trace mechanism in Equation (14) uses the coefficient ' δ ' as the ASE trace decay rate, which ensures that only the most recently visited states are eligible when updating the ASE weights.

5.3.3 - The Adaptive Critic Element (ACE)

The internal reinforcement signal ($\hat{r}$) is provided by the ACE, and is used to modify control actions and state evaluations continuously without having to wait until an external failure signal occurs. This is achieved on the basis of the negative reward signal (r) provided to the ACE. The reward signal is set to zero until a failure occurs, at which point it is set to -1. The ACE evaluates the performance of the cart-pole system and maintains its own set of evaluation weights (v_i). These are used to generate the internal reinforcement signal, which is then used to update the control weights (u_i) in the ASE. Additionally, the internal reinforcement signal is used by the ACE to update its own evaluation weights (v_i). Equation (15) shows how the internal reinforcement signal is calculated:

$$\hat{r}(t)=r(t)+\gamma\ \ p(t)-p(t-1) \tag{15}$$

The $\hat{r}$ signal is effectively the *temporal difference error* (TD error) described in Chapter Two, and is calculated by comparing the prediction of failure in the current state with the prediction of failure in the previous state. This error will be positive if the system moves from an "unsafe" to a "safe" state, and negative if it is the other way round. In

Equation (15), 'γ' is the discount factor, and $p(t)$ is the current prediction of expected failure. The discount factor determines how strongly predictions of failure are able to influence current actions, so that when $\gamma = 0$, then only the external reward signal $r(t)$ and the previous prediction $p(t-1)$ are taken into account. When γ is close to 1, actions are strongly influenced by the current prediction of failure. A more thorough explanation of the theory can be found in [Sutton, 1988]. Essentially, the prediction is the ACE weight (v_i) associated with the current state, given by Equation (16):

$$p(t) = v_i(t)d_i(t) \tag{16}$$

It is clear that since the negative reward signal (r) is either negative or zero, then the prediction of failure $p(t)$ will always be negative, and the ACE weight in Equation (16) will also be negative. A strong negative value for the weight (v_i) indicates that after visiting this state, the cart-pole system often failed. A value close to zero indicates that the state is associated with prolonged balancing periods. Hence Barto et al. [1983] described the weights as reflecting a prediction of failure, and called $p(t)$ the *prediction signal*. The external reward signal leads to punishment of all recent control actions leading to the failure, hence increases the prediction of failure in all recently entered states.

The ACE weights are updated continuously in accordance with the internal reinforcement signal ($\hat{r}$), and this is shown in Equation (17):

$$v_i(t+1) = v_i(t) + \beta \hat{r}(t)\bar{a}_i(t) \tag{17}$$

In Equation (17), 'β' is a positive learning constant, and $\bar{a}_i(t)$ is the eligibility trace for the ACE weights. This trace is calculated according to Equation (18):

$$\bar{a}_i(t+1)=\lambda\bar{a}_i(t)+(1-\lambda)d_i(t) \tag{18}$$

The trace mechanism in Equation (18) uses the coefficient 'λ' similar to TD(λ) described in Chapter Two. It is a method by which only the most recently visited states are eligible for update, and works in the same way as the ASE trace. This trace, however, has an effect on the *predictions* of future reinforcement produced by the ACE.

To summarise the performance of the ASE/ACE system, the ASE learns how to balance the pole in the absence of continuous reinforcement due to the internal reinforcement signal ($\hat{r}$) provided by the ACE. The internal reinforcement signal is computed by comparing the expectation of predicted failure (p) in current and previous states, and is used to update the ACE weights (v_i) in accordance with the recency information provided by the eligibility traces. The internal reinforcement signal is also used to update the ASE weights (u_i), which reflect appropriate control actions for every state such that the pole remains balanced for long periods of time.

5.4 - Simulated Pole Balancing: A Benchmark

The ASE/ACE system for the pole balancing problem has been simulated extensively by a number of researchers (e.g. Barto et al. [1983], Geva & Sitte [1992, 1993]). These simulations have shown that the ASE/ACE system is very robust on the pole balancing problem, but very slow on more complicated problems with large numbers of variables to be controlled [Barto et al., 1983]. In all simulations, the cart-pole system is always started from the centre of the track with the system variables initially set to zero. The small random noise term used in Equation (12) is sufficient for the control system to produce an initial action, and then the dynamical equations take over. The period that the

control system is able keep the pole balanced is called a 'trial'. A number of such trials (typically 100) constitutes a 'run'. Pole balancing experiments usually consist of a number of these runs, with the results averaged across runs.

An experimental benchmark specification to allow comparison between various approaches to solving the pole balancing problem was presented by Geva & Sitte [1993]. This followed their review of the literature on learning methods for cart-pole controllers which found that published results were difficult to compare because researchers did not keep to a uniform specification of the learning task, which also meant that there was no clear evidence that any one method was better than another. The benchmark proposed a set of standardised parameters to be used in pole balancing simulations so that all researchers could work on an equivalent problem. The benchmark parameters are given in Table 5.2.

System parameter	Value
Length of the track	± 2.4 m
Failure angles	± π/2 rad
Gravity (g)	-9.81 m/s^2
Length of the pole (2l)	1.0 m
Mass of the cart (m_c)	1.0 kg
Mass of the pole (m_p)	0.1 kg
Control force (F)	10.0 N
Integration time step	0.02 s

Table 5.2 - Benchmark System Parameters (Adapted from Geva & Sitte [1993])

The values of the learning coefficients used in the equations that make up the ASE/ACE system were also standardised, and these are given in Table 5.3.

Coefficient	Description	Value
α	ASE learning rate	1000.0
δ	ASE trace decay rate	0.9
γ	Discount factor	0.95
β	ACE learning rate	0.5
λ	ACE trace decay rate	0.8

Table 5.3 - Benchmark Coefficients (From Geva & Sitte [1993])

Geva & Sitte [1993] argued that the ability to balance the pole and centre the cart was not sufficient proof that effective learning had taken place, and that most learning methods designed to deal with the pole balancing problem were little better than a random search in parameter (weight) space. The benchmark specification therefore tries to make it very unlikely to find good controllers by chance, and proposes that simulations of the pole balancing problem should have a number of key features. These features are :-

1. *Demonstrate gradual improvements in system performance as individual learning sessions progress*
2. *Produce controllers that are able to centre the cart even when started away from the centre of the track with the pole already balanced*
3. *Use the standardised parameters specified by the benchmark*
4. *Use an unbounded track length and a failure angle of 90 degrees*
5. *Do not use friction because friction is difficult to determine, and so the performance of a simulated controller should not be affected by parameters hard to control in practice.*

Bapi et al. [1997] implemented the benchmark features described in experimental simulations of the pole balancing problem. Their work used a Neuro-Resistive Grid (NRG) as a replacement for the ACE, and their NRG simulation results were compared to simulations of the original ASE/ACE implementation. This work merits discussion because results from the NRG implementation have raised important questions regarding the relearning problem, and can be related to biological systems. Bapi et al. included all of the benchmark features described. Some changes to the original ASE/ACE representation scheme were made to allow states to have connections with their 'neighbours', and details of this change can be found in [Bapi et al., 1997]. The result is a grid of 162 nodes, where every node is connected to all its possible neighbours as shown in Figure 5.4.

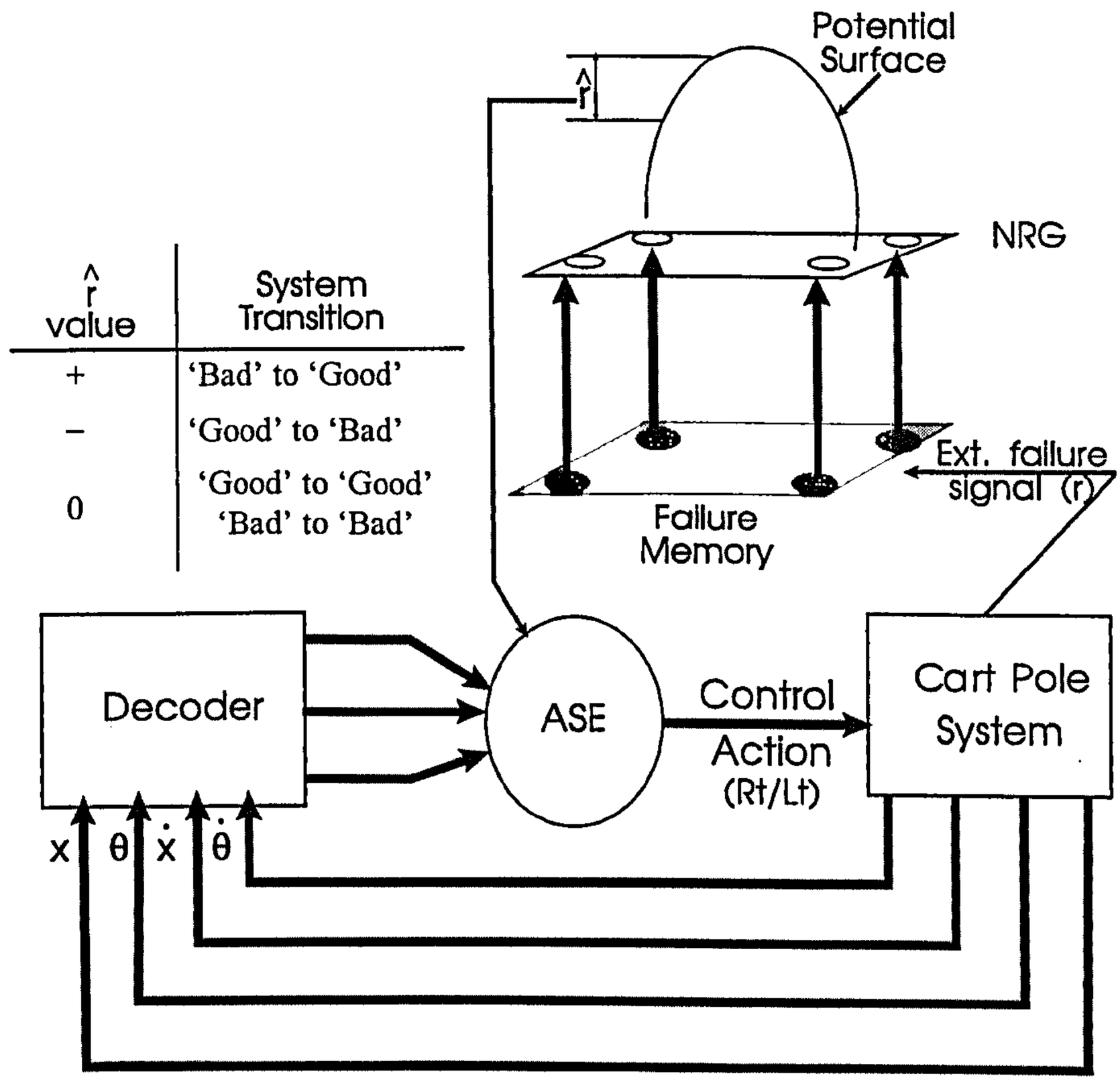


Figure 5.4 - The Neuro-Resistive Grid as ACE (From Bapi et al. [1997])

The NRG works like the ACE in that it receives a negative reward signal on failure, and is able to provide the ASE with an internal reinforcement signal ($\hat{r}$) by comparing the NRG value of the present state to the previous state. The NRG system is composed of two layers. The 'Failure Memory' layer records when a failure occurs in a particular state. This is equivalent to the role of the amygdala component as described in Chapter Four. The 'NRG' layer takes this information, and uses it to propagate failure information to neighbouring states in the grid. This means that the NRG learns more rapidly and with fewer computations than the original ASE/ACE system [Bapi et al., 1997]. With the NRG method however, if a failure occurs in a particular state then that state remains clamped as failed for the duration of a run. This is obviously not desirable if reinforcement contingencies change unexpectedly and relearning is required. The system will be unable to alter the value of a state even if it becomes part of a viable sequence of control actions in the future. Bapi et al. suggest that this problem deserves further analysis, and the work presented in this thesis contributes towards a better understanding of the problem.

5.5 - Relearning in the Pole Balancing Problem

The benefits of a simulation benchmark for pole balancing experiments have been described above. Unfortunately, the benchmark does not consider the relearning problem. For existing and future research to be compared, relearning needs to be introduced into the benchmark framework with a minimal disruption to the specifications that already exist. When investigating relearning, the objective is to change something in the system and then observe what happens to the performance of the system. A few studies have addressed related issues (such as robustness), but it is difficult to compare their work with this or any other work because of differences in the experiments conducted. These studies include :-

- Lin & Lin [1996] who used an ART-based fuzzy adaptive network called RFALCON to replace the ASE and ACE, and compared this to the original ASE/ACE system. Their experiments tested the disturbance rejection capability of the 'trained' system following a change to the system, e.g. changing the pole length or the cart mass. Their results suggested that the original ASE/ACE system needs additional trials to relearn. They did not use the benchmark so results cannot be directly compared.

- Santiago & Werbos [1995] changed the dimensions of the pole in its environment, and argued that this provides a challenging problem for many different types of controller. They also did not use the benchmark so results cannot be directly compared.

Both the above studies look at the robustness of the control system, and are thus different to relearning on the basis of reinforcement as defined in this work. The relearning problem in this work involves allowing the control system to learn about its environment, and then changing the reinforcement contingency rather than simply changing a system parameter. From Chapter Three, it is likely that the problem of dealing with relearning as a result of changes to system parameters will be the responsibility of the septo-hippocampal system. The amygdaloid complex is involved in dealing with relearning on the basis of changes to reinforcement, and how to achieve this with the simulation benchmark for pole balancing is the main objective of this chapter. This work therefore represents a contribution towards establishing a relearning benchmark for the pole balancing problem on the basis of changes to the reinforcement contingency. This involves specifying the experiments that could be conducted in order to investigate this particular relearning problem. A relearning benchmark must consider :-

- How can 'success' be defined in the pole balancing problem, i.e. when can it be said that the system has adequately learned or relearned to balance the pole?

- What experiments should be conducted so that the effect of changing the reinforcement contingency may be investigated using different approaches?

The standardised parameters specified in the original benchmark proposed by Geva and Sitte [1993] will be kept because these are not affected by extending the benchmark, and continue to allow comparison between new and existing control system designs. However, the criteria by which the system may be considered to have successfully learned to balance the pole needs to be established. Santiago & Werbos [1995] defined success as being able to balance the pole for 30 minutes. They stated that each "cycle" (time step) represented one second, therefore 30 minutes would equate to 1800 time steps. Simulation time depends on how long the computer takes to process each time step. The pole balancing benchmark specifies that 0.02 seconds represents each time step, therefore 30 minutes would represent 90,000 time steps in the benchmark. The availability of simulation time restricts the use of this criterion, and it is therefore proposed that 5000 successive time steps in any trial without failure is an adequate criterion to indicate 'success' has occurred. This criterion is better than the Santiago & Werbos [1995] criterion, and will allow performance to be compared over a specified number of time steps. However, because the criterion is used for comparison purposes only, then there is no guarantee that the system will not fail again after success. The aim is to observe the effect of changing the reinforcement contingency, i.e. the number of further trials and time steps it takes for the system to achieve success rather than to sustain success. Further trials would require additional reinforcements, and this could be considered as raising the costs

associated with relearning. Further time steps would extend the learning time, which is potentially a computational expense and again raises the costs associated with relearning. The time to relearn (in terms of the number of trials) starts from the trial immediately after a particular change is made, and thus the number of trials that it takes to initially achieve success is important. A number of experimental simulations were conducted to investigate the effect on learning and relearning in the original ASE/ACE system using the following features specified by the benchmark :-

- frictionless dynamical equations and standardised parameters
- a failure angle of 90 degrees and a failure length of 2.4m.

The remaining benchmark features were not included because these were designed to evaluate the *quality* of the eventual controller produced by the system once it has "learned" to balance the pole. Given that the criterion of 5000 time steps is not enough for the system to fully converge to a solution (Barto & Sutton [1983] used 100,000 time steps, and Geva & Sitte [1993] do not specify a criterion), then the rest of the benchmark is not appropriate to this work. The experimental method can be described as follows :-

- The experiments were simulated using Cortex-Pro, a DOS-based network simulation package with a BASIC-like language that provides many built-in functions for the simulation of nodes, networks, learning rules, transfer functions, etc., as well as a graphical user interface that allows access to all the network variables.
- Each experiment consisted of ten simulated runs, and the results from each run were tabulated. Each run is identified in the column headed '**Run**'. Although only the results from ten runs have been tabulated, these runs can be considered a sample from an

infinite number of potential runs. Ten runs provide sufficient data to allow an analysis of the performance of the ASE/ACE system in each of the relearning experiments, and this analysis is presented later in the chapter.

- Each run had a maximum of 200 trials for successful learning and relearning to occur. This was because the original Barto & Sutton [1983] simulations allowed one hundred trials for the system to 'learn' how to balance the pole, therefore an additional one hundred trials should provide sufficient opportunity for the system to 'relearn'.
- The system was allowed to achieve success, and the trial in which success occurred (i.e. the system managed to balance the pole for 5000 time steps without failure) was recorded, given in the column headed '**Learn**'. The cumulative time steps in all trials prior to the trial in which success occurred was recorded, given in the column headed '**Cum LTS**'. This was then averaged over all trials prior to the trial in which success occurred, given in the column headed '**Ave LTS**'. These measurements allowed the results to be averaged across all ten runs, thus giving an indicator of system performance for each experiment.
- After success, a change was made to the system taking effect in the trial immediately after the trial in which success was achieved. The system was again allowed to achieve success and the next trial in which success occurred was recorded, given in the column headed '**R-Trial**'.
- The column '**Relearn**' gives the difference between the first successful trial and the next successful trial, i.e. the number of intervening trials where the system was unsuccessful and needed additional negative reinforcement signals to enable it to relearn how to balance the pole.
- The cumulative time steps in all trials required to relearn was recorded, given in the column headed '**Cum RTS**'. This was then averaged over all relearn trials, given in the

column headed '**Ave RTS**'. These measurements were then averaged across all ten runs, thus giving an indicator of relearning performance for each experiment.

- Gaps in the column 'R-Trial' indicate that the system did not achieve success after the change, never managing to relearn before the 200 trials limit was reached.
- Gaps in the 'Relearn' column indicates that the system managed to achieve success in the very next trial so additional reinforcements were not required and this information is not appropriate. The justification for this is that additional reinforcement signals can be equated with extra costs, and therefore including runs with zero additional trials biases the overall evaluation of relearning performance.
- The row '**Mean**' gives the mean for each column, calculated by summing all the observed values in a column and then dividing this sum by the number of observations.
- The row '**St Dev**' gives the standard deviation for each column, calculated by taking the square root of the of the average of the squared distances of the observations from the mean. This gives a measure of the dispersion in the data.
- The row '**St Err**' gives the standard error of the mean for each column, calculated by dividing the standard deviation by the square root of the number of observations. This gives a measure of the precision in the data.
- Each run was terminated when the system managed to achieve success for a second time, or the number of trials exceeded 200.
- At the end of each run, all system variables and traces were reset. The traces were also reset at the end of each trial because this allowed the fastest learning speed for the adaptive critic system (see Bapi et al. [1997] for discussion).

5.6 - Performance of the ASE/ACE System on Relearning

The experimental results presented in the following sections consider the performance of the ASE/ACE system on the benchmark pole balancing problem extended to include relearning. In order to investigate relearning performance, we need to consider the performance of the system when something changes. The following experiments consider the effect of changing the pole length, the cart mass, and changing the failure length. There are numerous other experiments that would also have been appropriate (such as changing the failure angle), but these were not investigated.

5.6.1 - Changing the Pole Length

In this experiment, the benchmark ASE/ACE system was allowed to achieve success and then the length of the pole was changed by reducing it from 0.5m to 0.25m. This experiment is identical to the robustness test conducted by Lin & Lin [1996]. The results from ten runs (ASE/ACE and changing the pole length) are shown in Table 5.4.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9	14	1	4871	4871.0
2	8	900	128.6	11	2	4703	2351.5
3	20	13359	703.1	21			
4	11	3274	327.4	29	17	13887	816.9
5	45	14591	331.6	73	27	33606	1244.7
6	11	5709	570.9	12			
7	11	4751	475.1	13	1	599	599.0
8	18	4252	250.1	19			
9	22	19266	917.4	23			
10	23	14066	639.4	24			
Mean	18.1	9078.2	530.9		9.6	11533.2	1976.6
St Dev	10.8	6089.0	280.0		11.9	13260.3	1753.3
St Err	3.4	1925.5	88.6		5.3	5930.2	784.1

Table 5.4 - Time to Relearn: Change Pole Length (0.5m → 0.25m)

The results show that it took an average of 18.1 trials to learn to balance the pole, i.e. to balance the pole for 5000 time steps without failure. This gave a mean of 530.9 time steps per trial prior to success. The column 'Learn' shows the trial in which success was first achieved, and the length of the pole was changed at the start of the very next trial. It can be seen that five runs required additional trials (and therefore additional reinforcements) in order to 'relearn' to successfully balance the pole. This gives a mean of 9.6 trials and 1976.6 time steps per trial without success for these five runs.

5.6.2 - Changing the Cart Mass

In this experiment, the benchmark ASE/ACE system was allowed to achieve success and then the mass of the cart was changed by increasing it from 1.0kg to 2.0kg. This experiment is the same as the robustness test conducted by Lin & Lin [1996]. The results from ten runs (ASE/ACE and changing the cart mass) are shown in Table 5.5.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	2104	191.3	13			
2	19	3995	221.9	20			
3	14	12447	957.5	15			
4	18	3475	204.4	19			
5	12	8921	811.0	19	6	2188	364.7
6	16	2854	190.3	64	47	48876	1039.9
7	13	4059	338.3	14			
8	17	6809	425.6	18			
9	8	916	130.9	16	7	8932	1276.0
10	51	44829	896.6	53	1	4284	4284.0
Mean	18.0	9040.9	436.8		15.3	16070.0	1741.1
St Dev	12.1	13042.9	324.4		21.3	22051.5	1738.7
St Err	3.8	4124.5	102.6		10.7	11025.7	869.3

Table 5.5 - Time to Relearn: Change Cart Mass (1.0kg → 2.0kg)

The results show that it took a mean of 18.0 trials for the system to initially achieve success, which is consistent with the previous experiment. A mean of 436.8 time steps per trial were required. Four runs required trials to 'relearn', with a mean of 15.3 additional trials and 1741.1 time steps per trial for these runs. Figure 5.5 shows a screen dump taken during Run 1 of this simulation experiment. The screen is included to illustrate the simulation environment. The modulation graph shows that the 'λ' coefficient was a constant set to 0.8 throughout the experiment. The other coefficients were also set to benchmark values. It can be seen that the system first achieved success in trial 12, and the cart mass was changed at the start of the next trial. Success was achieved immediately after the change in trial 13 (the relearn trial), but the system failed again in trial 15. However, only the relearn trial is shown in Table 5.5 given the success criterion described earlier.

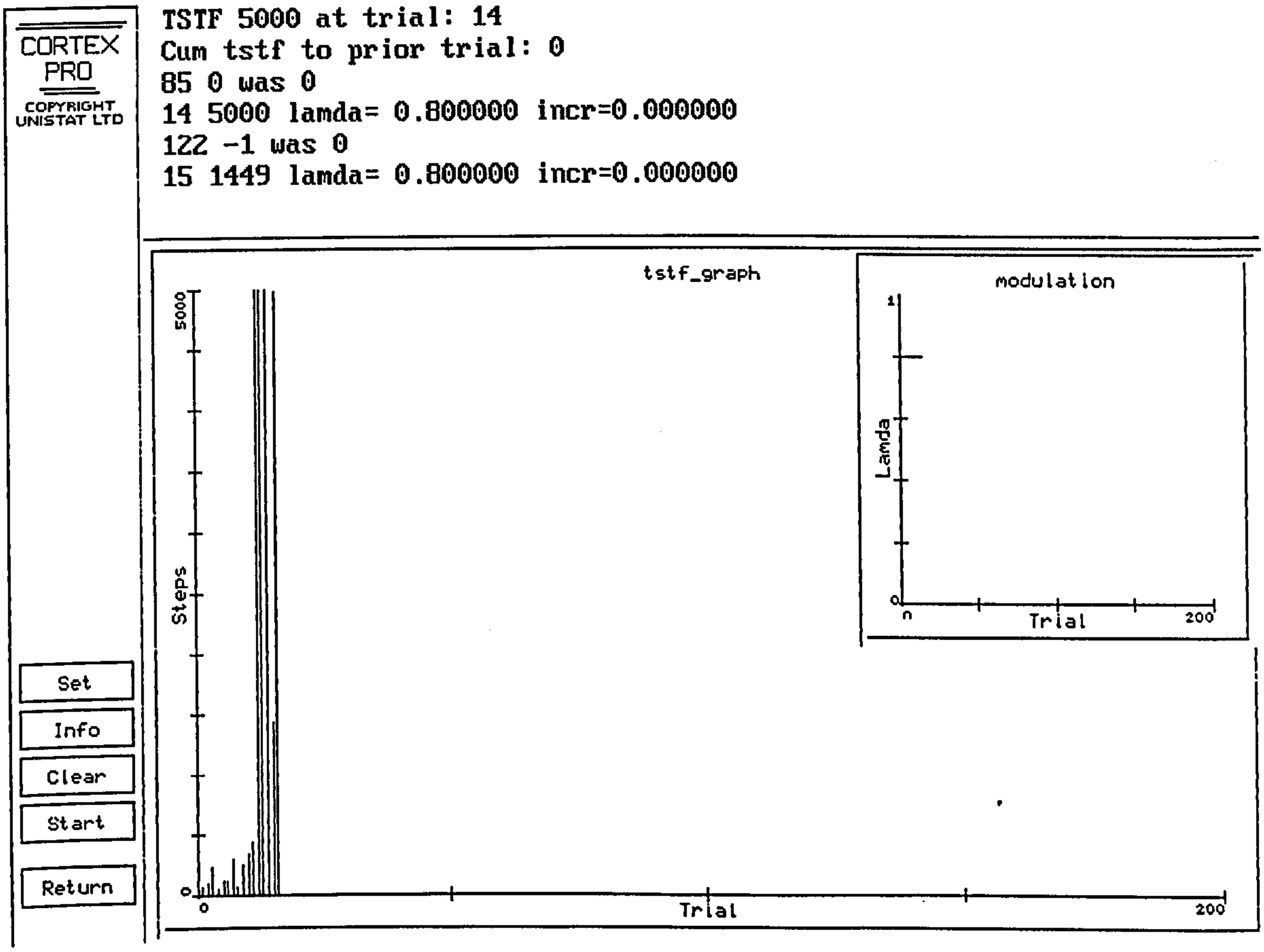


Figure 5.5 - Simulation Screen: Change Cart Mass (1.0kg → 2.0kg)

The first two experiments concur with the work of Lin & Lin [1996], and have been included because they do indeed suggest that the ASE/ACE system needs additional trials to 'relearn' once a change has been made to the system. These results also support the notion that balancing the pole for 5000 time steps is adequate for measuring learning performance because in all runs, the system managed to achieve this criterion in both learning and relearning phases despite a change to the system. This criterion will therefore allow comparison of relearning performance in subsequent experiments.

5.6.3 - Changing the Failure Length

Changing reinforcement contingencies is different from the previous experiments because the relationship between the state space and reinforcement signals remains the same, the only thing affected is how the controller directs the system into these states. The effect of changing reinforcement schedules needs to be investigated experimentally, and then related to the relearning problem. The next three experiments consider the effect of changing the failure length because this alters the position at which the system receives a negative reward signal. Changing the failure angle would also have been appropriate, but was not investigated. In the first experiment, the benchmark ASE/ACE system was allowed to achieve success, and then the failure length was changed from 2.4m to 1.5m. The results from ten runs are shown in Table 5.6.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	10	1641	182.3	11			
2	12	11821	1074.6	23	10	12065	1206.5
3	8	2709	387.0	9			
4	47	33810	735.0	58	10	20096	2009.6
5	16	5606	373.7	19	2	6226	3113.0
6	29	10460	373.6	35	5	1691	338.2
7	12	5241	476.5	19	6	2538	423.0
8							
9	141	94863	677.6	146	4	6288	1572.0
10	12	2704	245.8	20	7	9983	1426.1
Mean	31.9	18761.7	502.9		6.3	8412.4	1441.2
St Dev	42.8	30206.6	280.1		3.0	6341.6	952.5
St Err	14.3	10068.9	93.4		1.1	2396.9	360.0

Table 5.6 - Time to Relearn: Change Failure Length (2.4m → 1.5m)

The results show that it took a mean of 31.9 trials and 502.9 time steps per trial prior to success to learn how to balance the pole. This is inconsistent with the previous experiments, but it is likely that this mean reflects the result in Run 8 where the system never achieved success, and Run 9 where the system needed 141 time steps to initially achieve success. These runs could have been replaced by more successful runs, but this would give a false interpretation on the results in these experiments. There will always be extreme outliers, and it is the role of the standard error and similar statistics to identify their influence and thus reduce their significance in the interpretation of results. In this case, the standard error is low for relearning thus suggesting that relearning was not greatly affected. Seven runs actually required additional trials to 'relearn', with a mean of 6.3 trials and 1441.2 time steps per trial.

In the next experiment, the benchmark ASE/ACE system was allowed to achieve success, and then the failure length was changed from 2.4m to 1.0m. The results from ten runs are shown in Table 5.7.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9	42	29	17055	588.1
2	16	4577	305.1	100	83	70570	850.2
3	9	2364	295.5	52	42	38118	907.6
4	16	6344	422.9	30	13	2827	217.5
5	50	28693	585.6	62	11	10785	980.5
6	12	2415	219.5	43	30	11176	372.5
7	20	3373	177.5	36	15	5933	395.5
8	23	5560	252.7	40	16	7375	460.9
9	8	880	125.7	9			
10	11	4836	483.6	31	19	3077	161.9
Mean	17.7	6965.6	383.3		28.7	18546.2	548.3
St Dev	12.3	8095.4	248.8		22.7	22295.1	301.9
St Err	3.9	2560.0	78.7		7.6	7431.7	100.6

Table 5.7 - Time to Relearn: Change Failure Length (2.4m → 1.0m)

The results show that it took a mean of 17.7 trials and 383.3 time steps per trial prior to success to learn how to successfully balance the pole. All but one run required additional trials to 'relearn' to successfully balance the pole, with a mean of 28.7 trials and 548.3 time steps per trial.

In the third experiment, the benchmark ASE/ACE system was allowed to achieve success, and then the failure length was changed from 2.4m to 0.5m. The results from ten runs are shown in Table 5.8.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9				
2	14	7832	602.5				
3	21	9092	454.6	178	156	71203	456.4
4	9	3968	496.0	37	27	6048	224.0
5	10	1426	158.4	11			
6	8	1083	154.7				
7	13	3116	259.7				
8	14	2389	183.8	15			
9	27	6817	262.2				
10	13	2965	247.1	74	60	11633	193.9
Mean	14.1	4930.2	378.4		81.0	29628.0	291.4
St Dev	5.8	3386.6	257.1		67.0	36113.1	143.7
St Err	1.8	1070.9	81.3		38.7	20849.9	83.0

Table 5.8 - Time to Relearn: Change Failure Length (2.4m → 0.5m)

The results show that it took a mean of 14.1 trials and 378.4 time steps per trial prior to success to learn how to successfully balance the pole. Only five runs managed to 'relearn' to successfully balance the pole. Only three of these five runs required additional trials, with a mean of 81.0 trials and 291.4 time steps per trial. These results taken collectively suggest that the system had great difficulty in relearning.

All of the previous experiments considered the benchmark ASE/ACE system, and it would now be useful to compare the performance of the modified adaptive critic system (with amygdala component and modulated coefficients) with the results that have already been described. However, it is not possible to repeat all of the previous experiments using the modified adaptive critic system because of the simulation time that would be necessitated. Therefore, only one of these experiments will be selected as a suitable relearning experiment to allow comparisons between the benchmark ASE/ACE system and

the modified adaptive critic system. The mean number of trials and steps to learn and relearn for the benchmark ASE/ACE experiments is summarised in Table 5.9.

Exp.	Trials (L)	St Err	Steps (L)	St Err	N	Trials (R)	St Err	Steps (R)	St Err
Pole	18.1	3.4	530.9	88.6	5	9.6	5.3	1976.6	784.1
Mass	18.0	3.8	436.8	102.6	4	15.3	10.7	1741.1	869.3
Length = 1.5	31.9	14.3	502.9	93.4	7	6.3	1.1	1441.2	360.0
Length = 1.0	17.7	3.9	383.3	78.7	9	28.7	7.6	548.3	100.6
Length = 0.5	14.1	1.8	378.4	81.3	3	81.0	38.7	291.4	83.0

Table 5.9 - Summary of Results (Benchmark ASE/ACE System)

Table 5.9 shows the following information transposed from the results of the previous five experiments using the benchmark ASE/ACE system :-

'**Trials (L)**' - the mean number of trials to achieve success in the learning phase

'**Steps (L)**' - the mean number of steps required to achieve success in the learning phase

'**N**' - the number of runs that required additional trials in the relearning phase

'**Trials (R)**' - the mean number of additional trials required to achieve success in the relearning phase

'**Trials (R)**' - the mean number of steps per additional trial required in the relearning phase

'**St Err**' - the standard error for the value given in the preceding column

The results given in Table 5.9 suggest that changing the failure length from 2.4m to 1.0m is a suitable and challenging experiment for relearning. The change from 2.4m to 0.5m is clearly too difficult because in half of the runs relearning did not occur, and so this experiment is not suitable. The change from 2.4m to 1.5m could be used, but the change

from 2.4m to 1.0m is better because more runs show trials needing additional reinforcements to achieve relearning. This experiment is therefore a more challenging problem for relearning. It is now possible to investigate the modulation of coefficients in the modified adaptive critic system, and the effect that this has on relearning in the pole balancing problem.

5.7 - Performance of the Modified Adaptive Critic System on Relearning

It was suggested in Chapter Four that the adaptive critic system can be modified to include an 'amygdala component', and that the learning coefficients in the equations of the adaptive critic system can be modulated in accordance with information provided by the amygdala component. The modified adaptive critic system is shown again in Figure 5.6.

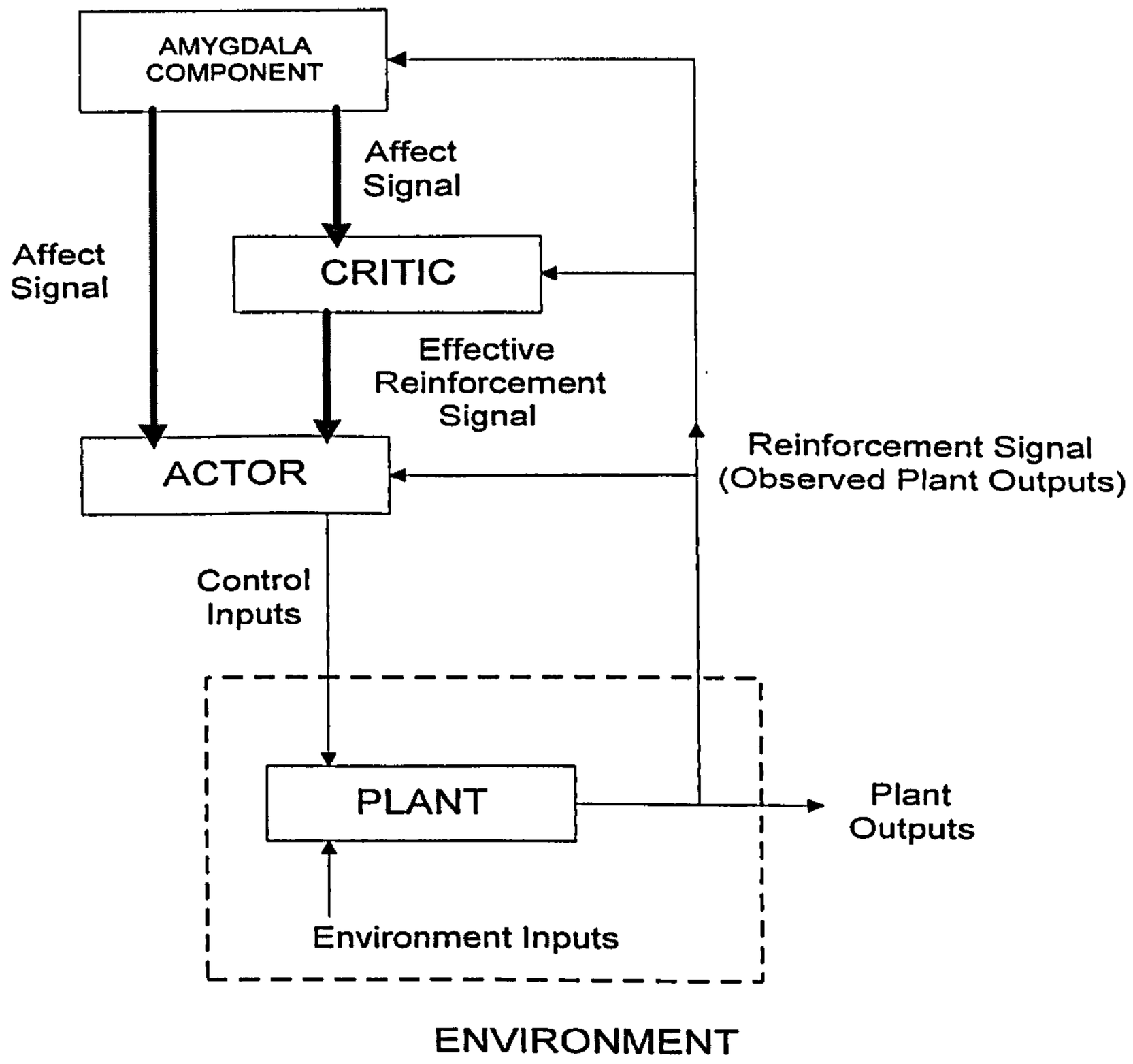


Figure 5.6 - Modified Adaptive Critic System

There are a number of possible experiments that can be conducted in order to investigate the effect of modulating coefficients in the learning and relearning phase of each experiment, as summarised in Table 5.10.

Learn	Relearn	Experiment
No Modulate	No Modulate	N/N
No Modulate	Modulate	N/M
Modulate	No Modulate	M/N
Modulate	Modulate	M/M

Table 5.10 - Modified Adaptive Critic Experiments

Table 5.10 shows that four possible experiments can be conducted to investigate the effect of modulating coefficients in the learning and relearning phase of each experiment. The first experiment (i.e. no modulation in either the learning or relearning phase) is equivalent to the experiments already conducted with the benchmark ASE/ACE system. The remaining three experiments will be conducted for each coefficient that is modulated, and the column 'Experiment' gives the legend that will be used when referring to these experiments.

It was also suggested in Chapter Four that the coefficients most likely to have an effect on relearning are the discount factor 'γ', and the temporal difference learning coefficient 'λ'. There are other coefficients in the adaptive critic system equations that affect learning which are also likely to influence relearning. The coefficient 'α' affects the learning rate of the actor (ASE) and the coefficient 'β' affects the learning rate of the critic (ACE). The hypothesis is that when a change in the reinforcement contingency is detected by the amygdala component, an affect signal will be provided that will enhance the effect

of a particular learning coefficient. The following sections will investigate this hypothesis experimentally using the modified adaptive critic system. Each experiment will use the relearning benchmark simulation parameters described previously, i.e. changing the failure length from 2.4m to 1.0m. The purpose of the amygdala component is to record when a reinforcement is received in a particular state, i.e. when the system fails in a particular state and the negative reward signal (r = -1) is provided. This was achieved by setting up an array of 162 elements, A, where each element corresponds to a system state. All elements in the array are initially set to zero. Whenever modulation of coefficients was required, the following rules were observed :-

If the system enters state x for the first time [A(x) = 0],

modulate coefficient because novelty (a new state) has been detected

let A{x) = -1 if the negative reward signal is received,

otherwise let A(x) = 1.

Else if the system enters a 'safe' state that it has previously encountered [A(x) = 1],

let A(x) = -1 if the negative reward signal is received, modulate coefficient,

else do not modulate coefficient, but allow it to decay.

Else if the system enters an 'unsafe' state that it has previously encountered [A(x) = -1]

let A(x) = 1 if the negative reward signal is not received, modulate coefficient

else do not modulate coefficient, but allow it to decay.

When the reinforcement contingency is changed as a result of changing the failure length from 2.4m to 1.0m, the above rules will allow any discrepancy between the expected reinforcement and the actual reinforcement to be detected. This discrepancy may be considered an 'unexpected reinforcement'. Note that this includes both situations

whereby a failure was expected in a state and not received, or a failure was received in a state but was not expected. In the benchmark relearning experiment defined here, only the latter could ever be true. This is because as the failure length is reduced, the number of states over which reinforcement signals can be received is increased, thus the system will never expect a reinforcement and not receive it. If the failure length was increased, then the former could be true and would also be considered an 'unexpected reinforcement'. This is therefore the basis for modulating the coefficients of interest. The coefficients 'γ', 'λ', 'α', and 'β' are modulated at the end of each trial by substituting them for *coefficient* in Equation (10):

$$coefficient_{trial} = \kappa * coefficient_{trial-1} + \omega \qquad (10)$$

The term 'κ' represents the decay rate, and the term 'ω' is an increment or decrement that will be described in each of the experiments. Every coefficient has a maximum and minimum value that has been arbitrarily set, and this provides a range between which each coefficient is modulated as described in the following experiments.

5.7.1 - Modulating the 'α' Learning Coefficient

The 'α' coefficient is the ASE learning coefficient given in Equation (13), and was a constant set to 1000.0 in the benchmark ASE/ACE system. The original Barto et al. simulations set the 'α' coefficient to a high value so that large changes would be reflected in the weights upon reinforcement [Barto et al., 1983]. This caused the probability of a rewarded action to become nearly one, and the probability of a punished action to become nearly zero. This was their attempt to make the system choose the same action each time its state was entered in any given trial, but Barto et al. conceded that this may be inappropriate

for problems where a particular state can receive conflicting reinforcement signals during trials. The relearning experiment used here is such a problem because states may receive different reinforcement signals when the failure length is changed. For the sake of comparison with the original ASE/ACE system, the 'α' coefficient will be modulated over a similar range of values in these experiments. The 'α' coefficient was initially set to a maximum value of 1000.0, and allowed to decay at the end of each trial to a minimum value of 500.0 with a decay rate of $\kappa = 0.8$. The increment $\omega = 200.0$ was added to the coefficient if the system failed in a state for the first time, or the system received an 'unexpected reinforcement'. The 'α' coefficient was thus modulated within the range $500.0 \leq \alpha \leq 1000.0$ during each run. The system was allowed to achieve success, and then the failure length was changed from 2.4m to 1.0m. The results from ten runs in all three 'α' coefficient modulation experiments (N/M, M/N and M/M) are shown in tables 5.11 to 5.13.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	15	2725	194.6	38	22	16333	441.4
2	8	1136	162.3	27	18	10686	411.0
3	23	8318	378.1	36	12	4589	131.1
4	23	7737	351.7	49	25	13428	279.8
5	57	21603	385.8	63	5	5934	95.7
6	11	7543	754.3	32	20	14440	465.8
7	32	11151	359.7	35	2	808	23.8
8	29	6959	248.5	45	15	10974	249.4
9	15	8496	606.9	27	11	11622	447.0
10	15	3696	264.0	16			
Mean	22.8	7936.4	370.6		14.4	9868.2	282.8
St Dev	14.3	5677.5	183.9		7.7	5071.6	168.9
St Err	4.5	1795.4	58.2		2.6	1690.5	56.3

Table 5.11 - Modulating the 'α' Coefficient, N/M

The results for the N/M experiment (Table 5.11) show that it took a mean of 22.8 trials and 370.6 time steps per trial to learn how to successfully balance the pole. Nine runs needed additional trials to 'relearn' with a mean of 14.4 trials and 282.8 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	17	6998	437.4	47	29	9979	216.9
2	13	2340	195.0	40	26	11632	298.3
3	20	6479	341.0	21			
4	15	2585	184.6	45	29	9197	209.0
5	12	3008	273.5	18	5	6346	373.3
6	12	3633	330.3	36	23	8796	251.3
7	13	3744	312.0	15	1	3170	226.4
8	15	4068	290.6	16			
9	13	6881	573.4	31	17	14892	496.4
10	13	3186	265.5	30	16	7308	252.0
Mean	14.3	4292.2	320.3		18.3	8915.0	290.5
St Dev	2.5	1800.9	114.8		10.6	3515.7	98.9
St Err	0.8	569.5	36.3		3.8	1243.0	35.0

Table 5.12 - Modulating the 'α' Coefficient, M/N

The results for the M/N experiment (Table 5.12) show that it took a mean of 14.3 trials and 320.3 time steps per trial to learn how to successfully balance the pole. Eight runs needed additional trials to 'relearn' with a mean of 18.3 trials and 290.5 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	42	23971	584.7	53	10	4491	449.1
2	16	3269	217.9	17			
3	30	6814	235.0	38	7	6814	973.4
4	16	6137	409.1	29	12	7779	648.3
5	49	41887	872.6	67	17	14581	857.7
6	37	17711	492.0	53	15	10061	670.7
7	11	3769	376.9	43	31	13101	422.6
8	12	2078	188.9	13			
9	19	4020	223.3	21	1	594	594.0
10	14	3018	232.2	15			
Mean	24.6	11267.4	383.3		13.3	8203.0	659.4
St Dev	13.8	12960.8	218.1		9.4	4863.2	201.1
St Err	4.4	4098.6	69.0		3.6	1838.1	76.0

Table 5.13 - Modulating the 'α' Coefficient, M/M

The results for the M/M experiment (Table 5.13) show that it took a mean of 24.6 trials and 383.3 time steps per trial to learn how to successfully balance the pole. Seven runs needed additional trials to 'relearn' with a mean of 13.3 trials and 659.4 time steps per trial.

5.7.2 - Modulating the 'β' Learning Coefficient

The 'β' coefficient is the ACE learning coefficient given in Equation (16), and was a constant set to 0.5 in the benchmark ASE/ACE system. In this experiment, the 'β' coefficient was initially set to 0.8, and allowed to decay at the end of each trial to a minimum value of 0.2 with a decay rate of $\kappa = 0.8$. The increment $\omega = 0.2$ was added to the coefficient each time the system encountered an unexpected reinforcement. The 'β' coefficient was thus modulated within the range $0.2 \leq \beta \leq 0.8$ as each run progressed. The results from ten runs in all three 'β' coefficient modulation experiments (N/M, M/N and M/M) are shown in tables 5.14 to 5.16.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	11	1333	133.3	16	4	3154	788.5
2	29	17001	607.2	44	14	3822	273.0
3	16	3208	213.9	18	1	4608	4608.0
4	21	8533	426.7	53	31	42341	1365.8
5	37	11426	317.4	39	1	53	53.0
6	13	2839	236.6	14			
7	17	3418	213.6	25	7	8343	1191.9
8	11	1675	167.5	28	16	16282	1017.6
9	56	14418	262.1	77	20	15123	756.2
10	10	1285	142.8	11			
Mean	22.1	6513.6	272.1		11.8	11715.8	1256.7
St Dev	14.8	5890.4	146.6		10.5	13648.1	1423.8
St Err	4.7	1862.7	46.3		3.7	4825.3	503.4

Table 5.14 - Modulating the 'β' Coefficient, N/M

The results for the N/M experiment (Table 5.14) show that it took a mean of 22.1 trials and 272.1 time steps per trial to learn how to successfully balance the pole. Eight runs needed additional trials to 'relearn' with a mean of 11.8 trials and 1256.7 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	35	8777	258.1	40	4	2043	510.8
2	99	39940	407.6	120	20	7568	378.4
3	10	1544	171.6	33	22	11698	531.7
4	13	3973	331.1	14			
5	10	1592	176.9	31	20	7436	371.8
6	13	4618	384.8	40	26	12490	480.4
7	12	4769	433.5	19	6	1906	317.7
8	36	10537	301.1	45	8	1859	232.4
9	25	7615	317.3	33	7	11206	1600.9
10	18	3474	204.4	19			
Mean	27.1	8683.9	298.6		14.1	7025.8	553.0
St Dev	27.1	11372.5	94.5		8.7	4588.2	435.4
St Err	8.6	3596.3	29.9		3.1	1622.2	153.9

Table 5.15 - Modulating the 'β' Coefficient, M/N

The results for the M/N experiment (Table 5.15) show that it took a mean of 27.1 trials and 298.6 time steps per trial to learn how to successfully balance the pole. Eight runs needed additional trials to 'relearn' with a mean of 14.1 trials and 553.0 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	29	10890	388.9	32	2	376	188.0
2	55	59137	1095.1	75	19	9257	487.2
3	13	4042	336.8	43	29	29836	1028.8
4	19	5579	309.9	24	4	6815	1703.8
5	12	2376	216.0	14	1	634	634.0
6	12	1516	137.8	13			
7	11	1813	181.3	14	2	297	148.5
8	25	3286	136.9	39	13	8720	670.8
9	7	799	133.2	8			
10	8	1427	203.9	20	11	6942	631.1
Mean	19.1	9086.5	314.0		10.1	7859.6	686.5
St Dev	14.5	17833.5	288.8		10.0	9658.4	498.0
St Err	4.6	5639.4	91.3		3.5	3414.8	176.1

Table 5.16 - Modulating the 'β' Coefficient, M/M

The results for the M/M experiment. (Table 5.16) show that it took a mean of 19.1 trials and 314.0 time steps per trial to learn how to successfully balance the pole. Eight runs needed additional trials to 'relearn' with a mean of 10.1 trials and 686.5 time steps per trial. Figure 5.7 shows a simulation screen taken during Run 2 of the M/M experiment.

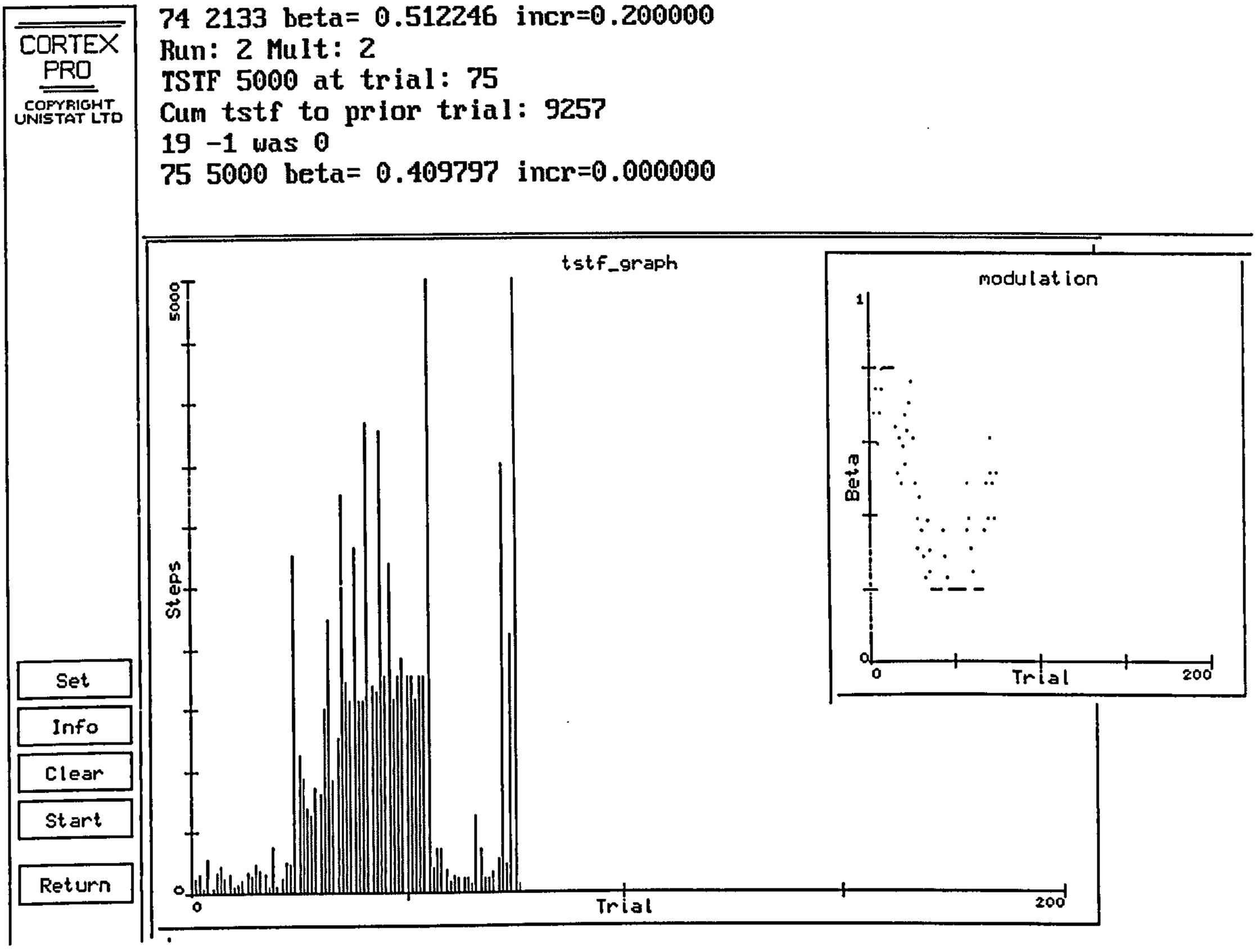


Figure 5.7 - Simulation Screen: Modulating 'β'

The screen shows that the 'β' coefficient was modulated between 0.2 and 0.8. All other coefficients were set to benchmark values. The system first achieved success in trial 55, and then the failure length was changed. Success was next achieved in trial 75.

5.7.3 - Modulating the 'γ' Discount Factor

The 'γ' coefficient is the discount factor given in Equation (15), and was a constant set to 0.95 in the benchmark ASE/ACE system. The closer the value of 'γ' is to one, the less the discount and hence the delayed consequence of the current prediction becomes more significant. For relearning, it is argued that unexpected reinforcements should enhance the learning of more recent reinforcements (described as a tactical objective in

Chapter Four), and hence 'γ' should be reduced to discount the effect of the current prediction in preference for the most recent reinforcement. The 'γ' coefficient was initially set to 0.5, and allowed to increase towards its maximum value of 0.95 at the end of each trial with a growth rate $\kappa = 1.25$. When an unexpected reinforcement occurred, the decrement $\omega = -0.5$ was added. This allowed the 'γ' coefficient to be modulated within the range $0.5 \leq \gamma \leq 0.95$ as each run progressed. Each time an unexpected reinforcement was detected, the value of 'γ' would go down thus increasing the discount of the current prediction, making this prediction less significant. The results from ten runs in all three 'γ' coefficient modulation experiments (N/M, M/N and M/M) are shown in tables 5.17 to 5.19.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9	35	22	15799	718.1
2	18	10041	590.6	47	28	16646	594.5
3	15	7586	541.9	20	4	4478	1119.5
4	13	3007	250.6	25	11	17433	1584.8
5	14	2416	185.8	25	10	13259	1325.9
6	13	3008	250.7	14			
7	10	1643	182.6	13	2	3108	1554.0
8	21	6180	309.0	26	4	3126	781.5
9	17	3827	239.2	71	53	39349	742.4
10	18	5074	298.5	31	12	3191	265.9
Mean	15.1	5339.6	381.4		16.2	12932.1	965.2
St Dev	3.3	3181.9	248.2		16.2	11722.1	454.9
St Err	1.1	1006.2	78.5		5.4	3907.4	151.6

Table 5.17 - Modulating the 'γ' Discount Factor, N/M

The results for the N/M experiment (Table 5.17) show that it took a mean of 15.1 trials and 381.4 time steps per trial to learn how to successfully balance the pole. Nine runs needed additional trials to 'relearn' with a mean of 16.2 trials and 965.2 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	125	42421	342.1	187	61	19054	312.4
2	124	89328	726.2	139	14	5924	423.1
3	133	115768	877.0	174	40	14397	359.9
4	177	72499	411.9	195	17	5009	294.6
5	26	13538	541.5	36	9	5817	646.3
6	74	14706	201.5	90	15	8990	599.3
7	72	72199	1016.9	81	8	5499	687.4
8	96	46351	487.9	130	33	19771	599.1
9	38	15908	429.9	62	23	8464	368.0
10	35	22946	674.9	60	24	8297	345.7
Mean	90.0	50566.4	571.0		24.4	10122.2	463.6
St Dev	49.7	35602.7	251.5		16.4	5599.5	151.7
St Err	15.7	11258.6	79.5		5.2	1770.7	48.0

Table 5.18 - Modulating the 'γ' Discount Factor, M/N

The results for the M/N experiment (Table 5.18) show that it took a mean of 90.0 trials and 571.0 time steps per trial to learn how to successfully balance the pole. The large mean number of trials to learn is interesting because the discount factor helps solve the temporal credit assignment problem by reducing the contribution of the current prediction, and therefore increases the contribution of previous predictions or actual reinforcements. Modulating the discount factor in the early stages of learning greatly decreases the contribution of the current prediction, and may therefore be detrimental to the initial learning process because the system loses the opportunity to acquire the necessary information in order to predict reinforcement. This may explain the high value for the mean number of trials to learn when using the modified adaptive critic system. All ten runs in this experiment needed additional trials to 'relearn' with a mean of 24.4 trials and 463.6 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	85	25418	302.6	93	7	6921	988.7
2	7	426	71.0	8			
3	69	44117	648.8	81	11	8784	798.5
4	54	9022	170.2	63	8	6923	865.4
5	52	18004	353.0	61	8	12025	1503.1
6	115	33467	293.6	160	44	25248	573.8
7	118	43442	371.3	159	40	34266	856.7
8	54	16639	313.9	73	18	5093	282.9
9	91	3166	35.2	135	43	13217	307.4
10	18	10519	618.8	20	1	476	476.0
Mean	66.3	20422.0	317.8		20.0	12550.3	739.2
St Dev	37.1	15749.1	202.1		17.3	10679.8	382.5
St Err	11.7	4980.3	63.9		5.8	3559.9	127.5

Table 5.19 - Modulating the 'γ' Discount Factor, M/M

The results for the M/M experiment (Table 5.19) show that it took a mean of 66.3 trials and 317.8 time steps per trial without success to learn how to successfully balance the pole. Table 5.19 also shows that nine runs needed additional trials to 'relearn' with an average of 20.0 trials and 739.2 time steps per trial. Figure 5.8 shows a simulation screen taken during Run 7 of the M/M experiment.

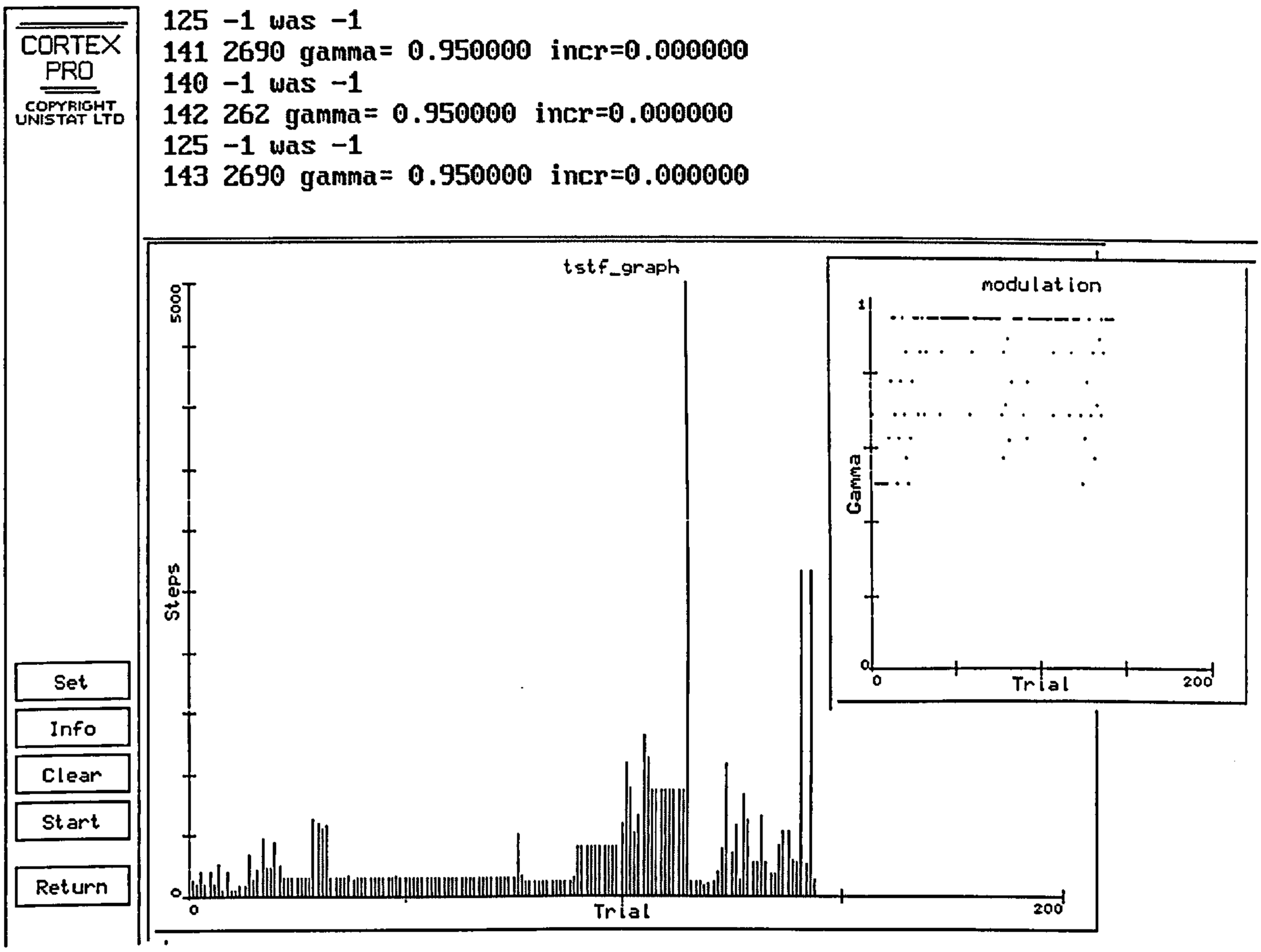


Figure 5.8 - Simulation Screen: Modulating 'γ'

Figure 5.8 shows that the system first achieved success in trial 118 of Run 7, and then the failure length was changed. Success was next achieved in trial 159 (not shown in the figure).

5.7.4 - Modulating the 'λ' Coefficient

The 'λ' coefficient is the coefficient used in the ACE eligibility trace given in Equation (17), and was a constant set to 0.8 in the benchmark ASE/ACE system. The closer the value of 'λ' is to one, the greater the number of past steps that will be considered and hence the greater the number of states that will eligible for update. For relearning, it may be argued that unexpected reinforcements should increase number of steps that are

considered, and hence 'λ' should also be increased when relearning is required. The 'λ' coefficient was initially set to 0.8 and allowed to decrease towards its minimum value of 0.2 at the end of each trial with a decay rate $\kappa = 0.8$. When an unexpected reinforcement occurred, the increment $\omega = 0.5$ was added. This allowed the 'λ' coefficient to be modulated within the range $0.2 \leq \lambda \leq 0.8$ as each run progressed. The results from ten runs in all three 'λ' coefficient modulation experiments (N/M, M/N and M/M) are shown in tables 5.20 to 5.22.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9	34	21	16604	790.7
2	8	3277	468.1	28	19	10069	529.9
3	18	7743	455.5	40	21	22341	1063.9
4	14	6947	534.4	15			
5	15	2605	186.1	35	19	21672	1140.6
6	23	7051	320.5	39	15	9127	608.5
7	43	32019	762.4	55	11	17368	1578.9
8	11	3963	396.3	12			
9	54	11090	209.2	55			
10	44	9249	215.1	56	11	7359	669.0
Mean	24.2	9455.8	451.2		16.7	14934.3	911.6
St Dev	16.5	8455.4	252.2		4.4	6106.7	372.4
St Err	5.2	2673.8	79.8		1.7	2308.1	140.7

Table 5.20 - Modulating the 'λ' Coefficient, N/M

The results for the N/M experiment (Table 5.20) show that it took a mean of 24.2 trials and 451.2 time steps per trial to learn how to successfully balance the pole. Seven runs needed additional trials to 'relearn' with a mean of 16.7 trials and 911.6 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	21	10614	530.7	47	25	15526	621.0
2	10	3277	364.1	11			
3	31	7743	258.1	44	12	11682	973.5
4	10	6947	771.9	11			
5	37	2605	72.4	72	34	25837	759.9
6	12	7051	641.0	14	1	1958	1958.0
7	19	32019	1778.8	37	17	7039	414.1
8	26	3963	158.5	27			
9	10	11090	1232.2	12	1	2887	2887.0
10	23	9249	420.4	59	35	11999	342.8
Mean	19.9	9455.8	622.8		17.9	10989.7	1136.6
St Dev	9.6	8455.4	526.9		14.2	8219.4	942.1
St Err	3.0	2673.8	166.6		5.4	3106.6	356.1

Table 5.21 - Modulating the 'λ' Coefficient, M/N

The results for the M/N experiment (Table 5.21) show that it took a mean of 19.9 trials and 622.8 time steps per trial to learn how to successfully balance the pole. Seven runs needed additional trials to 'relearn' with a mean of 17.9 trials and 1136.6 time steps per trial.

Run	Learn	Cum TSTF	Ave TSTF	R-Trial	Relearn	Cum TSTF	Ave TSTF
1	12	10614	964.9	25	12	7301	608.4
2	15	3730	266.4	27	11	7005	636.8
3	22	14583	694.4	32	9	11920	1324.4
4	20	8899	468.4	26	5	4648	929.6
5	20	10794	568.1	44	23	997	43.3
6	7	828	138.0	50	42	18969	451.6
7	28	10523	389.7	49	20	19595	979.8
8	17	4826	301.6	30	12	5369	447.4
9	14	6140	472.3	38	23	17411	757.0
10	18	12274	722.0	24	5	5538	1107.6
Mean	17.3	8321.1	498.6		16.2	9875.3	728.6
St Dev	5.8	4289.4	246.5		11.2	6652.2	372.9
St Err	1.8	1356.4	77.9		3.6	2103.6	117.9

Table 5.22 - Modulating the 'λ' Coefficient, M/M

The results for the M/M experiment show that it took a mean of 17.3 trials and 498.6 time steps per trial to learn how to successfully balance the pole. All ten runs required additional trials to 'relearn' with a mean of 16.2 trials and 728.6 time steps per trial. Figure 5.9 shows a simulation screen from Run 6 of the M/M experiment. The system first achieved success in trial 7, and then the failure length was changed. Success was next achieved in trial 50.

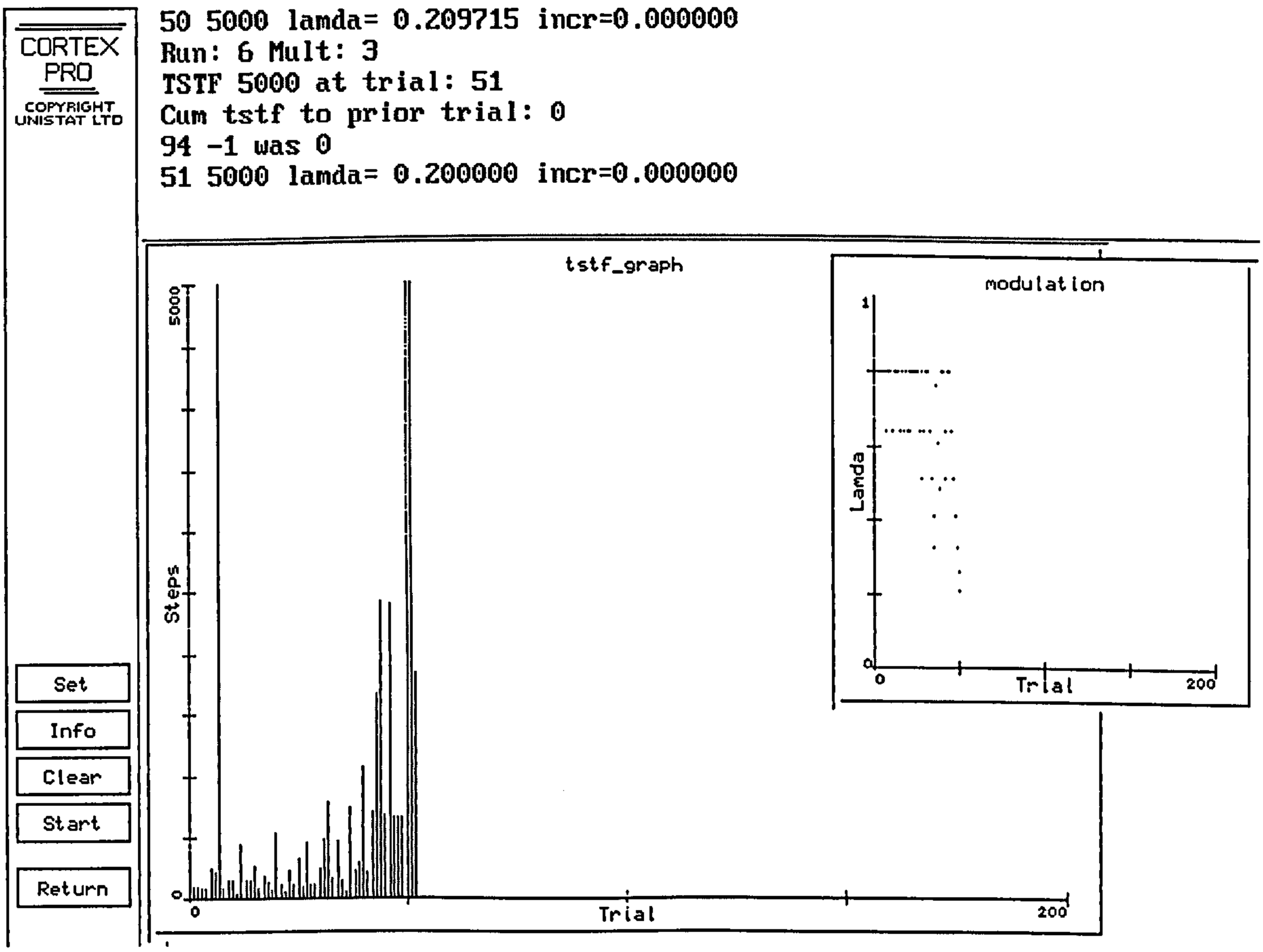


Figure 5.9 - Simulation Screen: Modulating 'λ'

5.8 - Analysis of Results

The previous section outlined a number of simulation experiments that were conducted to investigate the performance of the modified adaptive critic system with an amygdala component used as the basis for modulating coefficients. It is now appropriate to investigate the significance of the results from these experiments using statistical analysis. The results of the modified critic experiments, as well as the benchmark ASE/ACE system without modulation, are summarised in Table 5.23.

Exp.	Trials (L)	St Err	Steps (L)	St Err	n	Trials (R)	St Err	Steps (R)	St Err
No Mod	17.7	3.9	383.3	78.7	9	28.7	7.6	548.3	100.6
Alpha N/M	22.8	4.5	370.6	58.2	9	14.4	2.6	282.8	56.3
Alpha M/N	14.3	0.8	320.3	36.3	8	18.3	3.8	290.5	35.0
Alpha M/M	24.6	4.4	383.3	69.0	7	13.3	3.6	659.4	76.0
Beta N/M	22.1	4.7	272.1	46.3	8	11.8	3.7	1256.7	503.4
Beta M/N	27.1	8.6	298.6	29.9	8	14.1	3.1	553.0	153.9
Beta M/M	19.1	4.6	314.0	91.3	8	10.1	3.5	686.5	176.1
Gamma N/M	15.1	1.1	381.4	78.5	9	16.2	5.4	965.2	151.6
Gamma M/N	90.0	15.7	571.0	79.5	10	24.4	5.2	463.6	48.0
Gamma M/M	66.3	11.7	317.8	63.9	9	20.0	5.8	739.2	127.5
Lamda N/M	24.2	5.2	451.2	79.8	7	16.7	1.7	911.6	140.7
Lamda M/N	19.9	3.0	622.8	166.6	7	17.9	5.4	1136.6	356.1
Lamda M/M	17.3	1.8	498.6	77.9	10	16.2	3.6	728.6	117.9

Table 5.23 - Summary of Results (Modified Adaptive Critic System)

The following information has been transposed to Table 5.23 :-

'Trials (L)' - the mean number of trials to achieve success in the learning phase

'Steps (L)' - the mean number of steps required to achieve success in the learning phase

'n' - the number of runs that required additional trials in the relearning phase

'Trials (R)' - the mean number of additional trials required to achieve success in the relearning phase

'Trials (R)' - the mean number of steps per additional trial required in the relearning phase

'St Err' - the standard error for the value given in the preceding column

Each row in Table 5.23 represents an estimate of the mean trials and steps for learning and relearning (in each experiment) derived by taking a sample from the population as a whole. It is important to evaluate just how good these estimates are in case they are biased in some way. The most efficient estimates have the smallest standard errors and least variation in the data [Levin & Rubin, 1990]. As the sample size is less than 30 and the population standard deviation is unknown, we can use the *t* distribution to estimate the mean within a given confidence interval, which assumes that the population distribution is approximately normal. Table 5.24 gives the results for the ***learning phase***.

Exp.	Trials (L)	St Err	Limits	Steps (L)	St Err	Limits
No Mod	17.7	3.9	8.8	383.3	78.7	178.0
Alpha N	22.8	4.5	10.2	370.6	58.2	131.6
Alpha M	14.3	0.8	1.8	320.3	36.3	82.1
Alpha M	24.6	4.4	10.0	383.3	69.0	156.1
Beta N	22.1	4.7	10.6	272.1	46.3	104.7
Beta M	27.1	8.6	19.5	298.6	29.9	67.6
Beta M	19.1	4.6	10.4	314.0	91.3	206.5
Gamma N	15.1	1.1	2.5	381.4	78.5	177.6
Gamma M	90.0	15.7	35.5	571.0	79.5	179.8
Gamma M	66.3	11.7	26.5	317.8	63.9	144.5
Lamda N	24.2	5.2	11.8	451.2	79.8	180.5
Lamda M	19.9	3.0	6.8	622.8	166.6	376.8
Lamda M	17.3	1.8	4.1	498.6	77.9	176.2

Table 5.24 - Interval Estimates for Learn Phase (95% Confidence, $t = 2.262$)

In Table 5.24, the column 'Limits' determines the interval for the estimates at 95% confidence, and is calculated by multiplying the standard error by the *t* value (with 9 degrees of freedom). Each experiment indicates where modulation did (M) or did not (N) take place in the learning phase. For the benchmark ASE/ACE system (no modulation), the mean number of trials in the learning phase is in the interval 17.7 ± 8.8, and the mean number of steps in the learning phase is in the interval 383.3 ± 178.0, with 95% confidence. A similar interval can be calculated for all trials and steps using the formula 'Mean ± Limits'. The grey boxes indicate the mean results with the lowest standard errors used for hypothesis testing in Section 5.8.1. At the 95% confidence level, the results for the ***relearning phase*** are given in Table 5.25.

Exp.	n	t	Trials (R)	St Err	Limits	Steps (R)	St Err	Limits
No Mod	9	2.306	28.7	7.6	17.5	548.3	100.6	1763.1
Alpha M	9	2.306	14.4	2.6	6.0	282.8	56.3	337.6
Alpha N	8	2.365	18.3	3.8	9.0	290.5	35.0	314.5
Alpha M	7	2.447	13.3	3.6	8.8	659.4	76.0	669.5
Beta M	8	2.365	11.8	3.7	8.8	1256.7	503.4	4405.0
Beta N	8	2.365	14.1	3.1	7.3	553.0	153.9	1128.3
Beta M	8	2.365	10.1	3.5	8.3	686.5	176.1	1457.7
Gamma M	9	2.306	16.2	5.4	12.5	965.2	151.6	1887.8
Gamma N	10	2.262	24.4	5.2	11.8	463.6	48.0	564.6
Gamma M	9	2.306	20.0	5.8	13.4	739.2	127.5	1705.3
Lamda M	7	2.447	16.7	1.7	4.2	911.6	140.7	585.3
Lamda N	7	2.447	17.9	5.4	13.2	1136.6	356.1	4705.4
Lamda M	10	2.263	16.2	3.6	8.1	728.6	117.9	960.5

Table 5.25 - Interval Estimates for Relearn Phase (95% Confidence, *t* values shown)

In Table 5.25, the sample size varies between experiments and is given in the column headed 'n'. The column 'Limits' determines the interval for the estimates at 95%

confidence, and is calculated by multiplying the standard error by the *t* value (with n-1 degrees of freedom). Each experiment shows whether modulation did, M, or did not, N, take place in the relearning phase. For the benchmark ASE/ACE system (no modulation), the mean number of trials in the relearning phase is 28.7 ± 17.5, and the mean number of steps in the relearning phase is 548.3 ± 1763.1, with 95% confidence. A similar interval can be calculated for all trials and steps using the formula 'Mean ± Limits'. The grey boxes indicate the mean results with the lowest standard errors used for hypothesis testing in Section 5.8.1.

5.8.1 - Hypothesis Tests

The interpretation of these results is difficult given the relatively large interval observed for the benchmark ASE/ACE experiment, and the inconsistent nature of some of the results. There are some consistencies though, for example, the data suggests that modulating 'γ' leads to a greater number of trials in the learning phase, and thus a greater number of reinforcements are required. The data also suggests that modulating any of the coefficients investigated leads to a reduced number of trials in the relearning phase, and hence less reinforcements are required. This is accompanied by an increased number of steps when 'β' and 'γ' are modulated in the relearning phase. These findings can be tested statistically with one-tailed hypothesis tests for the difference between means. The null hypothesis is that there is no difference between means, the alternative hypotheses are :-

- that modulating 'γ' gives a higher mean number of trials in the learning phase than the mean without modulation (Test 1)
- that modulating the coefficients gives a lower mean number of trials in the relearning phase than the mean without modulation (Tests 2 - 5)

- that modulating the 'β' and 'γ' coefficients gives a greater number of steps in the relearning phase than the mean without modulation (Tests 6 and 7).

The alternative hypotheses are based on the difference between the two means, and thus these hypothesis tests will calculate the upper limit for acceptance of the null hypothesis for a given level of significance. Any result that falls outside this upper limit will cause the null hypothesis to be rejected in favour of the alternative hypothesis, which means that this result is statistically significant at given level of significance. This information is highlighted by the grey boxes in Table 5.26 which summarises all hypothesis tests, column headings are indicated below. The significance level will be set to 80% (the 95% limits are considered too restrictive).

Test	Mean1	Mean2	Diff	St Dev1	St Dev2	n1	n2	Pool Var	Sigma	$t_{80\%,n1+n2}$	Limit
1	66.3	15.1	51.2	37.1	3.3	10	10	693.7	11.8	0.861	10.1
2	18.3	14.4	3.9	10.6	7.7	8	9	84.1	4.5	0.863	3.8
3	14.1	10.1	4	8.7	10	8	8	87.8	4.7	0.865	4.1
4	24.4	16.2	8.2	16.4	16.2	10	9	265.9	7.5	0.861	6.5
5	739.2	463.6	275.6	382.5	151.7	9	10	81033.3	130.8	0.861	112.6
6	17.9	16.7	1.2	14.2	4.4	7	7	110.5	5.6	0.868	4.9
7	728.6	548.3	180.3	372.9	301.9	10	9	116508.2	156.8	0.861	135.0

Table 5.26 - Hypothesis Tests (80% Significance, *t* values shown)

- **'Test'** - the test number
- **'Mean1', 'Mean2'** - the two means tested (note that means selected are the ones with the lowest standard error)
- **'Diff'** - the difference between the means
- **'St Dev1', 'St Dev2'** - the respective standard deviation of the two means

- **'n1', 'n2'** - the size of sample from which the means are taken
- **'Pool Var'** - a pooled estimate of the population variance
- **'Sigma'** - the standard error of the difference between the two means
- '$t_{80\%,n1+n2}$' - the *t* value at 80% significance, and n1 + n2 degrees of freedom
- **'Limit'** - the upper limit for acceptance of the null hypothesis.

From these results it can be seen that at 80% significance levels, only Test 3 and Test 6 lead to accepting the null hypothesis i.e. may be interpreted as there being no difference between the means in these experiments. These results suggest that at 80% significance levels, there is no statistical evidence that modulating 'β' gives a lower mean number of trials in the relearning phase than the mean without modulation, nor that modulating 'λ' gives a lower mean number of trials in the relearning phase than the mean without modulation. The inference that can be drawn from this statistical analysis is that at 80% significance levels, there is a difference between the means in the other experiments conducted. This suggests the need for more detailed investigation of the modulation technique, and a greater understanding of the factors that have contributed to these findings. This is left to future research.

5.9 - Summary

This chapter described a benchmark pole balancing problem with standardised parameters. This benchmark was extended to allow consideration of the relearning problem, and a number of experimental simulations conducted to investigate the performance of the benchmark ACE/ASE system on relearning using the extended benchmark. The performance of the modified adaptive critic system using an amygdala

component and modulated coefficients on relearning was also investigated by experimental simulation, and the results from these simulations were presented in this chapter. Based on these experiments, there appear to be statistically significant results as regards the effect on relearning performance when using the modified adaptive critic system. This may be considered an opportunity for future work as discussed in the next chapter.

Chapter Six

Summary and Future Work

6.1 - Summary of the Thesis

The research conducted for this thesis looked at reinforcement learning and relearning in intelligent control, and considered how biological systems can provide inspiration for dealing with the relearning problem. This chapter will summarise the thesis, and outline the main contributions that it makes to knowledge. The limitations in this work will be considered, and the opportunities for future work discussed.

Chapter One provided a general introduction to the thesis, and described the various problems and requirements that have led to the need for this research. Intelligent control was described as a recent approach that aims to meet the demands of complex control problems. The relearning problem was described as one such problem that has resulted from the increasing demands made on control systems, such as the need for autonomous behaviour. It was argued that intelligent control attempts to address these problems by emulating the capabilities found in biological systems. These biological systems can be considered a useful source of inspiration for finding mechanisms that may perhaps be incorporated in intelligent control systems, and this is the rationale behind this research. The objectives of this work were therefore to investigate the reinforcement learning and relearning processes that occur in biological systems because these processes are likely to be involved in autonomous behaviour and other intelligent functions. This thesis specifically addressed relearning to see how biological systems deal with this problem. The aim was to see if there are any inherent mechanisms for dealing with

relearning found in biological systems that may provide inspiration for developing similar mechanisms in intelligent control systems.

Chapter Two provided a detailed look at learning control systems, and described how they are able to learn about their environment and adapt to changes in that environment. Reinforcement learning was presented as a framework within which learning control problems can be formulated, and the computational mechanisms behind reinforcement learning were discussed. These mechanisms can deal with many problems including the temporal credit assignment problem. The adaptive critic system was described as an approach that uses the reinforcement learning framework, and consists of five levels of adaptive critic design leading towards more "brain-like" control. These may ultimately meet the demands of intelligent control. The level one adaptive critic system uses a computational architecture that relies on a fixed schedule of reinforcement, and has been designed to learn about its environment with the assumption that the environment does not change. It does not possess the ability to detect changes in reinforcement schedules, although it can slowly deal with these changes (from results in Chapter Five). The relearning problem looks at how changes to reinforcement schedules can be detected, and how relearning can be achieved effectively, i.e. how to acquire new information at the same time as exploiting previous information that is still relevant. The thesis argued that the adaptive critic system can benefit from a different approach to address these issues, and needs to possess a mechanism that can more effectively deal with the relearning problem. This research therefore makes a contribution to knowledge by placing the relearning problem into the context of reinforcement learning, and considering how the adaptive critic system may be able to address this problem.

Chapter Three looked at reinforcement learning and relearning from a biological perspective, and proposed a conceptual model that describes how the amygdaloid complex

relearning found in biological systems that may provide inspiration for developing similar mechanisms in intelligent control systems.

Chapter Two provided a detailed look at learning control systems, and described how they are able to learn about their environment and adapt to changes in that environment. Reinforcement learning was presented as a framework within which learning control problems can be formulated, and the computational mechanisms behind reinforcement learning were discussed. These mechanisms can deal with many problems including the temporal credit assignment problem. The adaptive critic system was described as an approach that uses the reinforcement learning framework, and consists of five levels of adaptive critic design leading towards more "brain-like" control. These may ultimately meet the demands of intelligent control. The level one adaptive critic system uses a computational architecture that relies on a fixed schedule of reinforcement, and has been designed to learn about its environment with the assumption that the environment does not change. It does not possess the ability to detect changes in reinforcement schedules, although it can slowly deal with these changes (from results in Chapter Five). The relearning problem looks at how changes to reinforcement schedules can be detected, and how relearning can be achieved effectively, i.e. how to acquire new information at the same time as exploiting previous information that is still relevant. The thesis argued that the adaptive critic system can benefit from a different approach to address these issues, and needs to possess a mechanism that can more effectively deal with the relearning problem. This research therefore makes a contribution to knowledge by placing the relearning problem into the context of reinforcement learning, and considering how the adaptive critic system may be able to address this problem.

Chapter Three looked at reinforcement learning and relearning from a biological perspective, and proposed a conceptual model that describes how the amygdaloid complex

is involved in the processes of learning and relearning. The model proposes that the amygdaloid complex can influence a number of other systems when it detects that relearning is required, and this is achieved on the basis of detecting changes to the reinforcement contingencies. The model is based on the actions of various neurochemical substances that influence the amygdaloid complex and other structures. This conceptual model was an attempt to bring together evidence from many neurophysiological studies into a coherent model of how the amygdaloid complex is involved in reinforcement learning and relearning, and is thus a contribution to an understanding of the biological mechanisms involved. The model is speculative, but the development of such a model is a worthwhile endeavour consistent with the rationale behind this work, i.e. that biological mechanisms may provide inspiration for finding mechanisms to better address the relearning problem.

Chapter Four described the Houk et al. [1995b] model which proposes that structures in the basal ganglia are able to generate signals that predict reinforcement. This model has considerable structural overlap with the conceptual model from Chapter Three, as well as functional aspects that are remarkably similar to the adaptive critic system described in Chapter Two. The Houk et al. model does not consider the involvement of the amygdaloid complex as suggested by the conceptual model, and the chapter described how the adaptive critic system could be modified using the conceptual model to include an 'amygdala component'. This component records actual reinforcements, and is thus able to detect changes in the reinforcement contingency thereby influencing learning in both the actor and critic. These components are represented by the equations of the adaptive critic system. It was suggested that the influence of the amygdala component should be most significant when relearning is required (i.e. novelty or unexpected reinforcement), and that the learning coefficients in the adaptive critic equations can be modulated on this basis.

The thesis therefore makes a contribution to the computational modelling of biological mechanisms with regard to the amygdaloid complex and its involvement in relearning.

Chapter Five described pole balancing as a useful problem for investigating the behaviour of learning control systems, and presented a standardised benchmark problem specification. This benchmark was extended to allow consideration of the relearning problem by specifying possible experiments and parameters for investigating the relearning problem. A number of simulations were conducted to investigate the performance of the level one adaptive critic system on relearning using the extended benchmark. The performance of the modified adaptive critic system (with modulated coefficients) on the relearning problem was also investigated, and the results suggested statistically significant improvement in relearning performance with the modified adaptive critic system when compared to the performance of the level one adaptive critic system. This work represents a contribution to knowledge because it provides empirical results that can be used as the basis for future work on the relearning problem, and this may lead to further developments that may lead to addressing the relearning problem.

6.2 - Limitations and Future Work

The work in this thesis has a number of limitations. The discussion of the adaptive critic system in Chapter Two concentrates on the level one design, and is justified by the argument that biological mechanisms need to be understood on terms of the lower level designs before they can be applied to higher levels. A further justification is that some of the mechanisms used by the higher levels (such as backpropagation) are not biologically plausible given our current knowledge about the brain. It has been argued that higher level designs can already deal with some of the demands of complex control problems, such as

with Heuristic Dynamic Programming and its derivatives [Prokhorov & Wunsch II, 1996]. However, the higher level designs have no biological grounding, and do not look at the relearning problem explicitly. These higher level designs may benefit from looking at the problems of intelligent control (such as the relearning problem) from a biological perspective, and this is an opportunity for future work

The conceptual model presented in Chapter Three primarily considers the interaction between the amygdaloid complex and the basal ganglia, and does not account for the other structures and neurochemical systems involved in the same learning processes (although these are implicated). The role of other structures and neurochemical systems needs to be better understood and put into the context of relearning, such as the involvement of the septo-hippocampal system (e.g. Denham & McCabe [1996]) and neuropeptides (e.g. Gallagher [1984], Graybiel [1990]). This is an area that will gradually be addressed by ongoing and future research. Similarly, the conceptual model of the amygdaloid complex proposed by this thesis accounts for individual amygdaloid nuclei and their external connections to other structures. It does not consider the internal connections between the amygdaloid nuclei because these are not yet fully understood. The model is therefore based on assumption, and there is a need for future research to clarify how the intrinsic connections of the amygdaloid complex are involved in the processes of reinforcement learning and relearning, thus leading to a more complete model.

The modified adaptive critic system in Chapter Four is limited by the level of sophistication of the amygdala component, which is the basis for the modulation of learning coefficients. Modulation using the amygdala component needs to be refined so as to lead improved experimental results, and this is an area for future work. The modified adaptive critic system does not consider other biologically-inspired capabilities suggested by the conceptual model in Chapter Three. This includes the ability to switch between a

number of different control behaviours which may be potentially hard-wired into the system. This limitation arises from the fact that the experimental framework used is the benchmark pole balancing problem, which has only two possible control actions. Using a more difficult problem with a number of possible control actions or variable control strategies would enable the viability of a switching mechanism to be investigated, and is an excellent opportunity for future work.

The relearning experiments conducted in Chapter Five look at the effect of modulating individual coefficients on relearning. Based on simulation results, there appears to be an improvement in relearning performance in terms of the average number of trials needed to relearn with modulated learning coefficients. This modulation is equivalent to the interaction of neurochemical substances in biological systems, and is consistent with the activity in a number of different neural structures at the same time. This suggests that a number of learning coefficients will need to be modulated at the same time, and this needs to be investigated. This is an opportunity for future work and may lead to significant improvements in the performance of the system on relearning. The considerable activity that is now taking place towards understanding the underlying neurochemical systems will provide new knowledge, and this is likely to lead to a clearer understanding of which coefficients need to be modulated, when, and how. This is particularly useful for algorithms like TD(λ), where the value of the learning coefficient is specifically designed to influence the performance of the learning system. It would be interesting to investigate whether the coefficients have a *detrimental* influence on relearning performance when their modulation is counter to what would be expected, such as with 'γ' and 'λ' that have a strong theoretical basis. This aspect was not investigated, and represents an opportunity for future work.

It is important that the results presented in Chapter Five are placed into a wider context, as there has been very little work on the relearning problem from the perspective of changing reinforcement contingencies. The blocking and shortcut experiments with Dyna-Q [Sutton, 1990] use the changing worlds scenario which is similar, but integrates a number of approaches into a single algorithm. The objective is the same, i.e. how to exploit existing knowledge in order to reduce the costs of learning. Modulation of coefficients is equivalent to providing a 'boost' when relearning is necessary, very much in line with the exploration bonuses described by Dayan & Sejnowski [1996]. There are clearly advantages to using this technique, and what is now required is a better understanding of how to go about using parameter or coefficient modulation. The work in this thesis follows on from the work of Bapi et al. [1997], and the effect of modulated coefficients on relearning in the NRG system is an opportunity for future work. It is hoped that future research on the relearning problem will consider a variety of problems, and consider the modulation of coefficients in other reinforcement methods such as Q-Learning and advanced DP techniques. This will allow the results of this work to be put into a much broader context, which is necessary if the ideas presented by this thesis are to be realised in intelligent control systems.

References

[Aggleton, 1992] edited by: Aggleton, J.P.; The Amygdala: Neurobiological Aspects of Emotion, Memory and Mental Dysfunction, New York: Wiley-Liss.

[Alexander & Crutcher, 1990] Alexander, G.E.; Crutcher, M.D.; *Functional Architecture of Basal Ganglia Circuits: Neural Substrates of Parallel Processing*, TINS, vol.13, no.7, pp.266-271.

[Amaral et al., 1992] Amaral, D.G.; Price, J.L.; Pitkanen, A.; Carmichael, S.T.; *Anatomical Organisation of the Primate Amygdaloid Complex*, in [Aggleton, 1992], pp.1-66.

[Anderson, 1989] Anderson, C.W.; *Learning to Control an Inverted Pendulum Using Neural Networks*, IEEE Control Systems, April, pp.31-36.

[Antsaklis et al., 1991] Antsaklis, P.J.; Passino, K.P.; Wang, S.J.; *An Introduction to Autonomous Control Systems*, IEEE Control Systems, June, pp.5-13.

[Antsaklis et al., 1994] Antsaklis, P., (Chair); *Defining Intelligent Control: Report of the Task Force on Intelligent Control*, IEEE Control Systems, June, pp.4-5, 58-66.

[Antsaklis, 1995] Antsaklis, P.; *Intelligent Learning Control*, IEEE Control Systems, June, pp.5-9.

[Baker & Farrell, 1992] Baker, W.L.; Farrell, J.A.; *An Introduction to Connectionist Learning Control Systems*, in [White & Sofge, 1992], pp.35-63.

[Bapi et al., 1997] Bapi, R.S.; D'Cruz, B.; Bugmann, G.; *Neuro-Resistive Grid Approach to Trainable Controllers: A Pole Balancing Example*, Neural Computing & Applications, vol.5, pp.33-44.

[Barinaga, 1990] Barinaga, M.; *Amino Acids: How Much Excitement Is Too Much?*, Science, January, vol.247, pp.20-22.

[Barto et al., 1983] Barto, A.G.; Sutton, R.; Anderson, C.; *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*, IEEE Trans. Systems, Man, and Cybernetics, vol.13, no.5, pp.834-846.

[Barto, 1989] Barto, A.G.; *Connectonist Learning for Control: An Overview*, COINS Technical Report 89-89, University of Mass., pp.1-38.

[Barto et al., 1989] Barto, A.G.; Sutton, R.S.; Watkins, C.J.C.H.; *Learning and Sequential Decision-Making*, COINS Technical Report 89-95, University of Mass., pp.1-51.

[Barto, 1990] Barto, A.G.; *Connectionist Learning for Control*, in Miller III, W.T.; Sutton, R.S.; Werbos, P.J.; Neural Networks for Control, The MIT Press, pp.5-58.

[Barto, 1992] Barto, A.G.; *Reinforcement Learning and Adaptive Critic Methods*, in [White & Sofge, 1992], pp.469-491.

[Barto, 1995] Barto, A.G.; *Adaptive Critics and the Basal Ganglia*, in [Houk et al., 1995a], pp.215-232.

[Barto et al., 1995] Barto, A.G.; Bradtke, S.J.; Singh, S.P.; *Learning to Act Using Real-Time Dynamic Programming*, Artificial Intelligence, vol.72, pp.81-138.

[Bordi & LeDoux, 1992] Bordi, F.; LeDoux, J.; *Sensory Tuning beyond the Sensory System: An Initial Analysis of Auditory Response Properties of Neurons in the Lateral Amygdaloid Nucleus and Overlying Areas of the Striatum*, J.Neuroscience, vol.12, no.7, pp.2493-2503.

[Bower, 1993] Bower, J.M.; *The Modulation of Learning State in a Biological Associative Memory: An In Vitro, In Vivo, and In Computo Study of Object Recognition in Mammalian Olfactory Cortex*, AI Review, vol.7, no.5, pp.261-269.

[Bozarth, 1991] Bozarth, M.A.; *The Mesolimbic Dopamine System as a Model Reward System*, in [Willner & Scheel-Kruger, 1991], pp.301-330.

[Brodal, 1992] Brodal, P.; *The Cerebral Cortex and Limbic Structures*, The Central Nervous System: Structure and Function, New York: Oxford University Press, pp.381-397.

[Brooks, 1986] Brooks, V.B.; *How Does the Limbic System Assist Motor Learning? A Limbic Comparator Hypothesis*, Brain Behav. Evol., vol.29, pp.29-53.

[Brown & Harris, 1994] Brown, M.; Harris, C.; *An Introduction to Learning Modelling and Control*, Neurofuzzy Adaptive Modelling and Control, London: Prentice Hall International, pp.1-29.

[Bunsey & Strupp, 1995] Bunsey, M.D.; Strupp, B.J.; *Specific Effects of Idazoxan in a Distraction Task: Evidence that Endogenous Norepinephrine Plays a Role in Selective Attention in Rats*, Behavioural Neuroscience, vol.109, no.5, pp.903-911.

[Burns et al., 1996] Burns, L.H.; Annett, L.; Kelley, A.E.; Everitt, B.J.; Robbins, T.W.; *Effects of Lesions to Amygdala, Ventral Subiculum, Medial Prefrontal Cortex, and Nucleus Accumbens on the Reaction to Novelty: Implication for Limbic-Striatal Interactions*, Behavioural Neuroscience, vol.110, no.1, pp.60-73.

[Cador et al., 1989] Cador, M. Robbins, T.W.; Everitt, B.J.; *Involvement of the Amygdala in Stimulus-Reward Associations: Interaction with the Ventral Striatum*, Neuroscience, vol.30, no.1, pp.77-86.

[Carlson, 1986] Carlson, N.R.; Physiology of Behaviour, Boston: Allyn and Bacon Inc., pp.505-596, 629-636.

[Catania, 1970] Catania, A.C.; *Reinforcement Schedules and Psychophysical Judgements: a Study of Some Temporal Properties of Behaviour*, in Schoenfeld, W.N. (ed.); The Theory of Reinforcement Schedules, New York: Appleton-Century-Crofts, pp.1-42.

[Chevalier & Deniau, 1990] Chevalier, G.; Deniau, J.M.; *Disinhibition as a Basic Process in the Expression of Striatal Functions*, TINS, vol.13, no.7, pp.277-280.

[Cohen & Servan-Schreiber, 1992] Cohen, J.D.; Servan-Schreiber, D.; *Context, Cortex and Dopamine: A Connectionist Approach to Behavior and Biology in Schizophrenia*, Psychological Review, vol.99, no.1, pp.45-77.

[Cools et al., 1991] Cools, A.R.; Van Den Bos, R.; Ploeger, G.; Ellenbroek, B.A.; *Gating Function of Noradrenaline in the Ventral Striatum: Its role in Behavioural Responses to Environmental and Pharmacological Challenges*, in [Willner & Scheel-Kruger, 1991], pp.141-173.

[Dayan & Sejnowski, 1996] Dayan, P.; Sejnowski, T.J.; *Exploration Bonuses and Dual Control*, Machine Learning, vol.25, no.1, pp.5-22

[Denham, 1994] Denham, M.J.; *Learning to Control Intelligently*, Proc. IEE Control'94, March 1994, pp.1374-1378.

[Denham & McCabe, 1996] Denham, M.J.; McCabe, S.L.; *A Model of the Interactions between Prefrontal Cortex, Septum and the Hippocampal System in the Learning and Recall of Goal-Directed Sensory-Motor Behaviours*, Research Report NRG-96-01, Neurodynamics Research Group, University of Plymouth.

[Gallagher, 1984] Gallagher, M.; *Neurochemical Modulation of Memory: A Case for Opioid Peptides*, in [Squire & Butters, 1984], pp.579-587.

[Gallagher & Holland, 1994] Gallagher, M.; Holland, P.C.; *The Amygdala Complex: Multiple Roles in Learning and Attention*, Proc. Natl. Acad. Sci. (USA), December, vol.91, pp.11771-11776.

[Geva & Sitte, 1992] Geva S.; Sitte, J.; *Is the Broom-Balancer A Useful Test Case for Training Methods?*, Proc. IEEE Intl. Workshop on Emerging Technologies and Factory Automation, Melbourne, August, pp.283-289.

[Geva & Sitte, 1993] Geva, S.; Sitte, J.; *A Cartpole Experiment Benchmark for Trainable Controllers*, IEEE Control Systems, vol.13, no.5, pp.40-51.

[Goldman-Rakic & Selemon, 1990] Goldman-Rakic, P.S.; Selemon, L.D.; *New Frontiers in Basal Ganglia Research*, TINS, vol.13, no.7, pp.241-244.

[Gray, 1991] Gray, J.A.; *Neural Systems, Emotion and Personality*, in Madden, J. (ed.); Neurobiology of Learning, Emotion and Affect, New York: Raven Press, pp.273-305.

[Gray, 1995] Gray, J.A.; *The Contents of Consciousness: A Neurophysiological Conjecture*, Behavioural and Brain Sciences, vol.18, pp.659-722.

[Graybiel, 1990] Graybiel, A.; *Neurotransmitters and Neuromodulators in the Basal Ganglia*, TINS, vol.13, no.7, pp.244-254.

[Graybiel & Kimura, 1995] Graybiel, A.M.; Kimura, M.; *Adaptive Neural Networks in the Basal Ganglia*, see [Houk et al., 1995a], pp.103-116.

[Groenewegen et al., 1991] Groenewegen, H.J.; Berendse, H.W.; Meredith, G.E.; Haber, S.N.; Voorn, P.; Wolters, J.G.; Lohman, A.H.M.; *Functional Anatomy of the Ventral, Limbic System-Innervated Striatum,* in [Willner & Scheel-Kruger, 1991], pp.19-59.

[Grossberg, 1987] Grossberg, S.; *Competitive Learning: From Interactive Activation to Adaptive Resonance*, Cognitive Science, vol.11, pp.23-63.

[Grossberg & Merrill, 1992] Grossberg, S.; Merrill, J.W.L.; *A Neural Network Model of Adaptively Timed Reinforcement Learning and Hippocampal Dynamics*, Cognitive Brain Research, vol.1, pp.3-38.

[Groves et al., 1995] Groves, P.M.; Garcia-Munoz, M.; Linder, J.C.; Manley, M.S.; Martone, M.E.; Young, S.J.; *Elements of the Intrinsic Organisation and Information Processing in the Neostriatum*, in [Houk et al., 1995a], pp.51-96.

[Gullapalli et al., 1994] Gullapalli, V.; Franklin, J.A.; Benbrahim, H.; *Acquiring Robot Skills via Reinforcement Learning*, IEEE Control Systems, vol.14, no.1, pp.13-24.

[Gupta & Rao, 1994] Gupta, M.M.; Rao, H.D.; *Neuro-Control Systems: A Tutorial*, Neuro-Control Systems: Theory and Applications, IEEE Press, pp.1-10.

[Hasselmo, 1994] Hasselmo, M.E.; *Runaway Synaptic Modification in Models of Cortex: Implications for Alzheimer's Disease*, Neural Networks, vol.7, no.1, pp.13-40.

[Hestenes, 1992] Hestenes, D.; *A Neural Network Theory of Manic Depressive Illness*, edited by: Levine, D.S.; Leven, S.J.; Motivation, Emotion and Goal Direction in Neural Networks, New Jersey: Lawrence Erlbaum Associates, pp.209-257.

[Holland & Gallagher, 1993] Holland, P.C.; Gallagher, M.; *Amygdala Central Nucleus Lesions Disrupt Increments, But Not Decrements, in Conditioned Stimulus Processing*, Behavioural Neuroscience, vol.107, no.2, pp.246-253.

[Houk et al., 1995a] edited by: Houk, J.C.; Davis, J.L.; Beiser, D.G.; Models of Information Processing in the Basal Ganglia, London: MIT Press.

[Houk et al., 1995b] Houk, J.C.; Adams, J.L.; Barto, A.G.; *A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement*, in [Houk et al., 1995a], pp.249-270.

[Introini-Collison et al., 1996] Introini-Collison, I.; Dalmaz, C.; McGaugh, J.; *Amygdala β-Noradrenergic Influences on Memory Storage Involve Cholinergic Activation*, Neurobiology of Learning and Memory, vol.65, pp.57-64.

[Jackson & Houghton, 1995] Jackson, S.; Houghton, G.; *Sensorimotor Selection and the Basal Ganglia: A Neural Network Model*, in [Houk et al., 1995a], pp.337-367.

[Kaelbling et al., 1996] Kaelbling, L.P.; Littman, M.L.; Moore, A.W.; *Reinforcement Learning: A Survey*, Journal of Artificial Intelligence Research, vol.4, pp.237-285.

[Kapp et al., 1984] Kapp, B.S.; Pascoe, J.B.; Bixler, M.A.; *The Amygdala: A Neuroanatomical Systems Approach to Its Contribution to Aversive Conditioning*, in [Squire & Butters, 1984], pp.473-488.

[Kapp et al., 1994] Kapp, B.S.; Supple, W.F.; Whalen, P.J.; *Effects of Electrical Stimulation of the Amygdaloid Central Nucleus on Neocortical Arousal in the Rabbit*, Behavioural Neuroscience, vol.108, no.1, pp.81-93.

[Kesner, 1992] Kesner, R.P.; *Learning and Memory in Rats With an Emphasis on the Role of the Amygdala*, in [Aggleton, 1992], pp.379-399.

[Kesner & Williams, 1995] Kesner, R.P.; Williams, J.M.; *Memory for Magnitude of Reinforcement: Dissociation Between the Amygdala and Hippocampus*, Neurobiology of Learning and Memory, vol.64, pp.237-244.

[Koob, 1992] Koob, G.F.; *Dopamine, Addiction and Reward*, Seminars in the Neurosciences, vol.4, pp.139-148.

[Larcombe, 1996] Larcombe, P.; *The Inverted Pendulum: Obscurity and Ubiquity*, Mathematics Today, Jan/Feb, pp.14-16.

[LeDoux et al., 1988] LeDoux, J.E.; Iwata, J.; Cicchetti, P.; Reis, D.J.; *Different Projections of the Central Amygdaloid Nucleus Mediate Autonomic and Behavioural Correlates of Conditioned Fear*, J.Neuroscience, vol.8, no.7, pp.2517-2529.

[LeDoux, 1989] LeDoux, J.E.; *Cognitive-Emotional Interactions in the Brain*, Cognition and Emotion, vol.3, no.4, pp.267-289.

[LeDoux et al., 1990] LeDoux, J.E.; Cicchetti, P.; Xagoraris, A.; Romanski, L.M.; *The Lateral Amygdaloid Nucleus: Sensory Interface of the Amygdala in Fear Conditioning*, J.Neuroscience, vol.10, no.4, pp.1062-1069.

[LeDoux, 1992] LeDoux, J.E.; *Brain Mechanisms of Emotion and Emotional Learning*, Current Opinion in Neurobiology, vol.2, pp.191-197.

[LeDoux, 1995] LeDoux, J.E.; *Emotion: Clues from the Brain*, Annual Review of Psychology, vol.46, pp.209-305.

[LeDoux, 1997] LeDoux, J.E.; The Emotional Brain: The Mysterious Underpinnings of Emotional Life, New York: Simon & Schuster, pp.138-178, 267-303.

[Levin & Rubin, 1990] Levin, R.I.; Rubin, D.S.; Statistics for Management, New Jersey: Prentice Hall, pp.62-130, 256-413.

[Liang & McGaugh, 1983] Liang, K.C.; McGaugh, J.L.; *Lesions of the Stria Terminalis Attenuate the Amnestic Effect of Amygdaloid Stimulation on Avoidance Responses*, Brain Research, vol.274, pp.309-318.

[Lin & Lin, 1996] Lin, C-J; Lin, C-T; *Reinforcement Learning for an ART-Based Fuzzy Adaptive Learning Control Network*, IEEE Trans. Neural Networks, vol.7, no.3, pp.709-731.

[Llinas, 1990] Llinas, R.R.; *The Workings of the Brain: Development, Memory and Perception*, Editors Preface to "Readings from Scientific American", pp.i-xii.

[Maren & Fanselow, 1996] Maren, S.; Fanselow, M.S.; *The Amygdala and Fear Conditioning: Has the Nut Been Cracked ?*, Neuron, vol.16, pp.237-240.

[McRae-Degueurce et al., 1985] McRae-Degueurce, A.; Dennis, T.; Leger, L.; Scatton, B.; *Regulation of Noradrenergic Neuronal Activity in the Rat Locus Coeruleus by Serotoninergic Afferents*, Physiological Psychology, vol.13, no.3, pp.188-196.

[McDonald & White, 1993] McDonald, R.J.; White, N.M.; *A Triple Dissociation of Memory Systems: Hippocampus, Amygdala and Dorsal Striatum*, Behavioural Neuroscience, vol.107, no.1, pp.3-22.

[McGaugh et al., 1990] McGaugh, J.L.; Introini-Collison, I.B.; Nagahara, A.H.; Cahill, L.; *Involvement of the Amygdaloid Complex in Neuromodulatory Influences on Memory Storage*, Neuroscience and Biobehavioral Reviews, vol.14, pp.425-431.

[Millington & Baker, 1990] Millington, P.J.; Baker, W.L.; *Associative Reinforcement Learning for Optimal Control*, Proc. AIAA Guidance Navigation and Control'90, vol.2, pp.1120-1128.

[Mogenson & Yim, 1991] Mogenson, G.J.; Yim, C.C.; *Neuromodulatory Functions of the Mesolimbic Dopamine System: Electrophysiological and Behavioural Studies*, in [Willner & Scheel-Kruger, 1991], pp.105-130.

[Moore & Atkeson, 1993] Moore, A.W.; Atkeson, C.G.; *Prioritised Sweeping: Reinforcement Learning with Less Data and Less Real Time*, Machine Learning, vol.13, pp.103-130.

[Myers et al., 1996] Myers, C.E.; Ermita, B.R.; Harris, K.; Hasselmo, M.; Solomon, P.; Gluck, M.A.; *A Computational Modewl of Septohippocampal Activity in Classical Eyeblink Conditioning*, Neurobiology of Learning and Memory, vol.66, pp.51-66.

[Narendra, 1994] Narendra, K.S.; *Neural Networks for the Intelligent Control of Dynamical Systems*, Proc. WCNN'94, vol.2, pp.3-8.

[Oades, 1985] Oades, R.D.; *The Role of Noradrenaline in Tuning and Dopamine in Switching Between Signals in the CNS*, Neuroscience & Biobehavioral Reviews, vol. 9, pp.261-282.

[Ogata, 1990] Ogata, K.; *Adaptive Control Systems*, Modern Control Engineering, London: Prentice-Hall International, pp.850-856.

[Passino, 1993] Passino, K.M.; *Bridging the Gap Between Conventional and Intelligent Control*, IEEE Control Systems, June, pp.12-18.

[Posner & Peterson, 1990] Posner, M.I.; Peterson, S.E.; (1990), *The Attention System of the Human Brain*, Annual Review of Neuroscience, vol.13, pp.25-42.

[Prokhorov et al., 1995] Prokhorov, D.V.; Santiago, R.A.; Wunsch II, D.C.; *Adaptive Critic Designs: A Case Study for Neurocontrol*, Neural Networks, vol.8, no.9, pp.1367-1372.

[Prokhorov & Wunsch II, 1996] Prokhorov, D.V.; Wunsch II, D.C.; *Advanced Adaptive Critic Designs*, Proc. WCNN'96, pp.83-87.

[Ribeiro, 1995] Ribeiro, C.H.C.; *Attentional Mechanisms as a Strategy for Generalisation in the Q-Learning Algorithm*, Proc. ICANN'95, pp.455-460.

[Robbins et al., 1985] Robbins, T.W.; Everitt, B.J.; Cole, B.J.; Archer, T.; Mohammed, A.; *Functional Hypotheses of The Coeruleocortical Noradrenergic Projection: A Review of Recent Experimentation and Theory*, Physiological Psychology, vol.13, no.3, pp.127-150.

[Robbins, 1992] Robbins, T.W.; *Introduction: Milestones in Dopamine Research*, Seminars in The Neurosciences, vol.4, pp.93-97.

[Robbins & Everitt, 1992] Robbins, T.W.; Everitt, B.J.; *Functions of Dopamine in the Dorsal and Ventral Striatum*, Seminars in The Neurosciences, vol.4, pp.119-127.

[Rolls, 1990] Rolls, E.T.; *A Theory of Emotion and its Application to Understanding the Neural Basis of Emotion*, Cognition and Emotion, vol.4. no.3, pp.161-190.

[Routtenberg, 1980] edited by: Routtenberg, A.; Biology of Reinforcement: Facets of Brain-Stimulation Reward, New York: Academic Press.

[Rumelhart, 1997] Rumelhart, D.E.; *Affect and Neuro-modulation: A Connectionist Approach*, in Cohen, J.D.; Shuler, J.W.; (eds), Scientific Approaches to Consciousness, Mawah, New Jersey: Erlbaum, pp.469-477.

[Santiago & Werbos, 1995] Santiago, R.A.; Werbos, P.J.; *New Progess Towards Truly Brain-Like Intelligent Control*, Proc. WCNN'95, vol.1, pp.27-33.

[Sara, 1985] Sara, S.J.; *The Locus Coeruleus and Cognitive Function: Attempts to Relate Noradrenergic Enhancement of Signal/Noise in the Brain to Behaviour*, Physiological Psychology, vol.13, no.3, pp.151-162.

[Sara, 1988] Sara, S.J.; *Noradrenaline and Memory: Neuromodulatory Influences on Retrieval*, edited by Weinman, J.; Hunter, J.; (1991), Memory: Neurochemical and Abnormal Perspectives, London: Harwood Academic, pp.105-128.

[Sara et al., 1995] Sara, S.J.; Dyon-Laurent, C.; Herve, A.; *Novelty Seeking Behavior in the Rat is Dependent on the Integrity of the Noradrenergic System*, Cognitive Brain Research, vol.2, pp.181-187.

[Sarter & Markowitsch, 1985] Sarter, M.; Markowitsch, H.J.; *Involvement of the Amygdala in Learning and Memory: a Critical Review With Emphasis on Anatomical Relations*, Behavioural Neuroscience, vol.99, no.2, pp.342-380.

[Schultz et. al, 1995] Schultz, W.; Romo, R.; Ljungberg, T.; Mirenowicz, J.; Hollerman, J.R.; Dickinson, A.; *Reward-related Signals Carried by Dopamine Neurons*, see [Houk et al., 1995a], pp.233-248.

[Schwartz & Robbins, 1995] Schwartz, B.; Robbins, S.J.; Psychology of Learning and Behavior, New York: W. W. Norton & Company, pp.172-173.

[Segal, 1985] Segal, M.; *Mechanisms of Action of Noradrenaline in the Brain*, Physiological Psychology, vol.13, no.3, pp.172-178.

[Shoureshi, 1991] *Learning and Decision Making for Intelligent Control Systems*, IEEE Control Systems, January, pp.34-37.

[Squire & Butters, 1984] edited by: Squire, L.; Butters, N.; Neuropsychology of Memory, New York: The Guilford Press.

[Stein, 1980] Stein, L.; *The Chemistry of Reward*, in [Routtenberg, 1980], pp.109-130.

[Stellar & Stellar, 1985] Stellar, J.R.; Stellar, E.; *Behavioural Concepts and Definitions*, The Neurobiology of Motivation and Reward, New York: Springer-Verlag Inc., pp.25-50.

[Sutton & Barto, 1981] Sutton, R.; Barto, A.G.; *Toward a Modern Theory of Adaptive Networks: Expectation and Prediction*, Psychological Review, vol.88, no.2, pp.135-170.

[Sutton, 1988] Sutton, R.S.; *Leaning to Predict by the Methods of Temporal Differences*, Machine Learning, vol.3, pp.9-44.

[Sutton, 1990] Sutton, R.S.; *Integrated Architectures for Learning, Planning and reacting Based on Approximating Dynamic Programming*, Proc. Seventh Int'l Conf. on Machine Learning, San Mateo. CA: Morgan Kaufmann, pp.216-224.

[Sutton et al., 1992] Sutton, R.S.; Barto, A.G.; Williams, R.J.; *Reinforcement Learning is Direct Adaptive Optimal Control*, IEEE Control Systems, April, pp.19-22.

[Sutton & Barto, 1995] Sutton, R.S.; Barto, A.G.; *Reinforcement Learning I, III: Learning to Act/Applications*, Neural Networks Summer School (Course Notes), July, Cambridge University, pp.1-20.

[Swerdlow & Koob, 1987] Swerdlow, N.R.; Koob, G.F.; *Dopamine, Schizophrenia, Mania and Depression: Toward a Unified Hypothesis of Cortico-Striato-Pallido-Thalamic Function*, Behavioural and Brain Sciences, vol.10, pp.197-245.

[Tesauro, 1995] Tesauro, G.; *Temporal Difference Learning and TD-Gammon*, Communications of the ACM, vol.38, no.3, pp.58-67.

[Thrun, 1992] Thrun, S.B.; *The Role of Exploration in Learning Control*, in [White & Sofge, 1992], pp.527-560.

[Watkins, 1989] Watkins, C.J.C.H.; Learning from Delayed Rewards, Ph.D. Thesis, University of Cambridge.

[Weiner, 1990] Weiner, I.; *Neural Substrates of Latent Inhibition*, Psychological Bulletin, vol.108, no.3, pp.442-461.

[Werbos, 1991] Werbos, P.; *An Overview of Neural Networks for Control*, IEEE Control Systems, January, pp.40-41.

[Werbos, 1994] Werbos, P.; *Neural Nets, Consciousness, Ethics and the Soul*, Proc. WCNN'94, San Diego, vol.1, pp.221-228.

[Werbos, 1995] Werbos, P.; *Optimal Neurocontrol: Practical Benefits, New Results and Biological Evidence*, Proc. WCNN'95, Washington D.C., vol.2, pp.318-325.

[White & Sofge, 1992] edited by: White, D.A.; Sofge, D.A.; Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches, New York: Van Rostrand Reinhold.

[Wickens & Kotter, 1995] Wickens, J.; Kotter, R.; *Cellular Models of Reinforcement*, in [Houk et al., 1995a], pp.187-214.

[Widrow et al., 1973] Widrow, B.; Gupta, N.K.; Maitra, S.; *Punish/Reward: Learning with a Critic in Adaptive Threshold Systems*, IEEE Trans. Systems, Man, and Cybernetics, vol.3, pp.455-465.

[Wieland, 1991] Wieland, A.P.; *Evolving Neural Network Controllers for Unstable Systems*, Proc. IJCNN'91, vol.2, pp.667-673.

[Willner, 1983] Willner, P.; *Dopamine and Depression: A Review of Recent Evidence*, Brain Research Reviews, vol.6, pp.225-236.

[Willner & Scheel-Kruger, 1991] edited by: Willner, P.; Scheel-Kruger, J.; The Mesolimbic Dopamine System: From Motivation to Action, London: John Wiley & Sons.

Neural Comput & Applic (1997)5:33–44

Neural
Computing
& Applications

Neuro-Resistive Grid Approach to Trainable Controllers: A Pole Balancing Example

Raju S. Bapi, Brendan D'Cruz and Guido Bugmann

Neurodynamics Research Group, School of Computing, University of Plymouth, Plymouth, UK

A new neural network approach is described for the task of pole-balancing, considered a benchmark learning control problem. This approach combines Barto, Sutton and Anderson's [1] Associative Search Element (ASE) with a Neuro-Resistive Grid (NRG) [2] acting as Adaptive Critic Element (ACE). The novel feature in NRG is that it provides evaluation of a state based on propagation of the failure information to the neighbours in the grid. NRG is updated only on a failure, and provides ASE with a continuous internal reinforcement signal by comparing the value of the present state to the previous state. The resulting system learns more rapidly and with fewer computations than that of Barto et al. *[1]. To establish a uniform basis of comparison of algorithms for pole balancing, both the systems are simulated using benchmark parameters and tests specified in Geva and Sitte [3].*

Keywords: Inverted pendulum problem; Reinforcement learning; Learning control; Nonlinear control; Neural networks; Neuro-resistive grid method; Value map

1. Introduction

Pole-balancing or balancing-an-inverted-pendulum has been identified as a benchmark problem for trainable controllers [3]. The task consists of balancing a pole, attached vertically to a movable cart, by applying one dimensional forces of constant magnitude to the base of the cart. The controller does not have access to the equations of motion of the system. The general problem here is to discover a sequence of binary (right or left) control forces that can keep the system balanced for long periods of time. To enable this discovery, the only information available to the learning control system is a negative reinforcement signal given when the system collapses. Hence, the controller faces the problem of evaluating its intermediate actions in the absence of any continuous external information. The strategy in these delayed reinforcement problems is to select the actions at every step that would reduce the possibility of eventual failure.

The system state space is determined by four variables, namely, the position of the cart on the rail (measured with reference to the centre of the railing), linear velocity of the cart, the angle of inclination of the pole (measured with reference to the vertical line), and the angular velocity of the pole. To discover a sequence of correct control actions over this four dimensional space, it is easier if the space is quantised so that the problem remains tractable. The quantisation can be fixed *a priori* or learnt from the system behaviour [4]. In the solution proposed in this paper, as was the case in [1] also, we will assume that the quantisation is determined *a priori*. The effects of such a fixed quantisation scheme on the performance of both the algorithms are discussed at the end of Sect. 4. The system equations, parameters, and the state space quantisation scheme are shown in Table 1. The controller is deemed to have failed if the pole angle exceeds the specified limit or the cart reaches either end point of the one dimensional railing. Based on the failure signal the system needs to adjust its decisions and internal mechanisms that give rise to these decisions such that the future performance is improved.

Several algorithms have been proposed for solving

Correspondence and offprint requests to: R. Bapi, Neurodynamics Research Group, School of Computing, University of Plymouth, Plymouth PL4 8AA, UK. email: rajubapi@soc.plym.ac.uk

Table 1. (a) System parameters: (b) state space quantisation scheme; (c) polecart system equations (a minor error in the system equations given in Geva and Sitte [3] has been corrected); (d) diagram of the polecart system.

Polecart parameter	Value
Length of the track	2.4 m
Failure angles (θ)	$\pm$ $\pi/2$ rad
Gravity (g)	-9.81 m/s^2
Length of the pole (2l)	1 m
Mass of the cart (m_c)	1 kg
Mass of the pole (m_p)	0.1 kg
Control force (F)	10 N
Integration time step	0.02 s

a

Variable	Range	Region
x (m)	$[-2.4,-0.8)$	1
	$[-0.8,0.8]$	2
	$(0.8,2.4]$	3
θ (rad)	$[-1.57,-0.21)$	1
	$[-0.21,-0.02)$	2
	$[-0.02,0.00)$	3
	$[0.00,0.02]$	4
	$(0.02,0.21]$	5
	$(0.21,1.57]$	6
$\dot{x}$ (m/s)	$(-\infty,-0.5)$	1
	$[-0.5,0.5]$	2
	$(0.5,+\infty)$	3
$\dot{\theta}$ (rad/s)	$(-\infty,-0.87)$	1
	$[-0.87,0.87]$	2
	$(0.87,+\infty)$	3

b

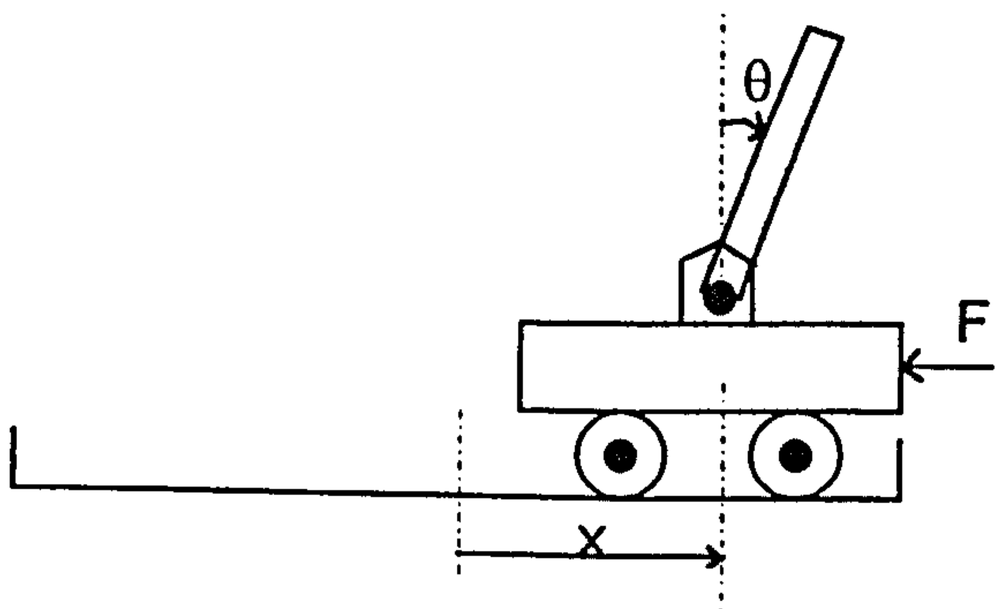
Table 1. continued

$$\frac{d^2\theta}{dt^2} = \frac{g\sin\theta - a\cos\theta - \mu_p \dot{\theta}^2 l \cos\theta \sin\theta}{l\left(\frac{4}{3} - \mu_p \cos^2\theta\right)}$$

where $a = \dfrac{F}{m_p + m_c}$ and $\mu_p = \dfrac{m_p}{m_p + m_c}$

$$\frac{d^2x}{dt^2} = \frac{\frac{4}{3}a + \left(\frac{4}{3}\dot{\theta}^2 l - g\cos\theta\right)\mu_p \sin\theta}{\left(\frac{4}{3} - \mu_p \cos^2\theta\right)}$$

c



d

this problem (see Geva and Sitte [3] for a good review), but ASE/ACE method appears the most general method of all these. The adaptive critic approach belongs to the family of Temporal Difference (TD) learning algorithms wherein system states are evaluated based on a delayed reinforcement signal [5]. The work presented here reports simulation results of the ASE/ACE system of Barto *et al.* [1] on benchmark parameters, and compares this approach with the new method using the neuro-resistive grid.

In the adaptive critic method, Barto *et al.* [1] introduced a way of coping with delayed reinforcement information. Each system state is associated with an action and its evaluation. The action associated with a state is stored in the value of the weight connecting the decoder and the Associative Search Element (ASE). The evaluation of a state is stored in the value of the weight connecting the decoder and the Adaptive Critic Element (ACE). Value attached to a state indicates how often that state was part of a sequence of actions that led to system collapse. The system has to enter a state before its evaluation can be initiated. So it takes a long time to build the value map over the state space, and thereby leads to long learning times with this algorithm. In the work reported here, the basic architecture of Barto *et al.* [1] has been retained except for replacing the Adaptive Critic Element (ACE) with the Neuro-Resistive Grid (NRG).

The design of the NRG technique was inspired

by the shape of the evaluation maps produced by the Temporal Difference (TD) learning methods in a maze solving task [6]. The potential distribution in the resistive grid results from the flow of current away from the goal state. It does not exactly have the same shape as the evaluation map, and hence it does not produce exactly the same sequence of actions as in the TD learning method. However, qualitatively the resistive grid technique captures the essence of TD methods, assigning higher values to points close to the goal and proposing realisable paths that avoid obstacles. Due to the lateral connections in the resistive grid, evaluations spread rapidly which improves generalisation, which is a weak point of TD learning methods [6]. TD learning methods are part of dynamic programming methods such as Q-learning which evaluate state-action pairs instead of states alone [7]. For Q-learning also, it was observed that lateral spread of evaluations leads to better generalisation [8]. Lateral communication also reduces exploration time and accelerates learning, as will be outlined in later sections.

A detailed description of the adaptive critic approach is given in the next section. Neuro-resistive grid method is outlined in Sect. 3, and its application to the pole-balancing problem is given in Sect. 4. Simulation results of these two systems (adaptive critic approach and the neuro-resistive grid method) are presented with a comparative discussion in Sect. 5. We will conclude with a summary and an outline of future directions in Sect. 6.

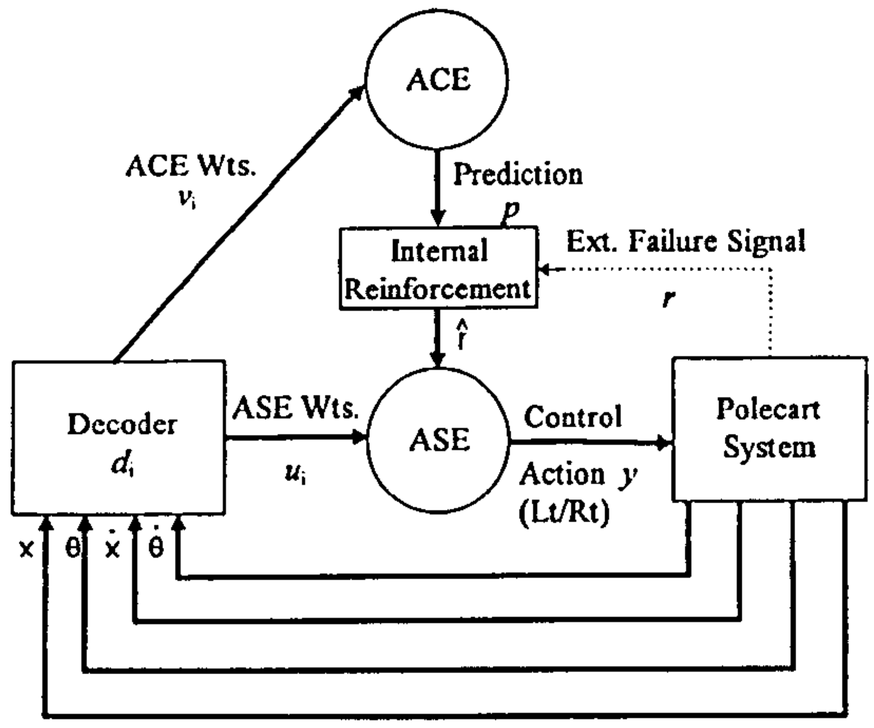
2. Adaptive Critic Approach

Since the new method proposed here is a variation of the adaptive critic approach of Barto *et al.* [1], the latter will be described in detail below. The control system configuration of Barto *et al.* [1] is shown in Table 2(a) and the equations for the system components are shown in Table 2(b). In the controller proposed by them, input to the controller is a vector of four variables specifying current values of the cart position (x), its velocity ($\dot{x}$), angle of the pole (θ), and its angular velocity ($\dot{\theta}$). This four dimensional space is then quantised (the quantisation parameters are given in Table 1) into 162 regions by the decoder. The decoder converts the input state vector into a 162-bit binary number (d_i) that has a unit value for the bit corresponding to the region that the input vector belongs to and a zero value for the remaining bits. Each of the regions (boxes) of the decoder is connected to both the Associative Search Element (ASE) and the Adaptive Critic Element (ACE) by a weight, u_i and v_i, respectively. In other words, each of the 162 system states has its own ASE and ACE weights, u_i and v_i, respectively. Based on the sign of the weight (u_i), ASE generates a binary control action (y), which in turn is transformed into a control force (F) to be applied to the right or left of the base of the cart. The weights (v_i) attached to the Adaptive Critic Element (ACE) reflect the evaluation of states of the system. The ACE receives an external reinforcement signal (r) when the pole cart system fails. The temporal traces, e_i and $\bar{a}_i$, in ASE and ACE, respectively, help keep track of the time elapsed from the last visit to the state. This trace mechanism helps apportioning the blame for failure of the system to various states. As there is no intermediate reinforcement signal available to evaluate the actions of the ASE, an internal reinforcement signal ($\hat{r}$) is generated using the ACE weights (v_i). The $\hat{r}$ signal is used to train the ASE weights.

The weights to ASE and ACE are set to zero initially. The pole cart system always starts from the centre of the railing with zero initial condition on all the variables. Initial control actions from the ASE are random because of the noise term ($noise(t)$) in the ASE output equation in Table 2(b). This is a Gaussian noise term with zero mean and a standard deviation of 0.01. This small random term enables the controller to explore the space in the absence of a known control action. The learning in ASE results in enhancing the weights, so that eventually the weight value can overcome the random noise term, thereby generating a known (non-random) control action. The control action amounts merely to applying a constant control force (F) on either the left or the right of the base of the cart. The system equations (in Table 1(c)) are updated using this control force. The equations are integrated numerically using Euler's method with a time step of 0.02 s. The decoder determines the next box that the system enters by decoding the new system state vector. The above process is continued until the system fails. The time period that the system keeps the pole balanced is called a 'trial'. In our simulations, one hundred such trials constitute a 'run'. A run is terminated before 100 trials if the total time of balance exceeds 500 000 time steps as in Barto *et al.* [1]. This leads to unequal number of trials in different runs. To compare the performance across 'runs', the data has to be adapted suitably. These details are discussed in Sect. 5.

On the very first trial in each run, there is no evaluation ($p(t)$) from ACE as the initial ACE weights (v_i) are set to zero. The external failure signal (r) is always set to zero and made equal to -1 only when the system fails. So when a failure

Table 2. (a) ASE/ACE system configuration. Based on the state parameters from the polecart system, decoder determines the box (d_i) that the system entered. This information is used to determine the control action (y) and update all the weights (u_i and v_i) (See text for more details.); (b) equations for the adaptive critic system.



a

ACE EQUATIONS:
Prediction from ACE:

$$p(t) = v_i(t)d_i(t);\ d_i(t) \text{ is the decoded system state}$$

ACE weights:

$$v_i(t+1) = v_i(t) + \beta\, \hat{r}(t)\bar{a}_i(t);\ \beta \text{ is a positive learning constant}$$

ACE eligibility traces:

$$\bar{a}_i(t+1) = \lambda \bar{a}_i(t) + (1-\lambda)d_i(t);\ \lambda \text{ is trace decay rate constant}$$

Internal reinforcement signal:

$$\hat{r}(t) = r(t) + \gamma\, p(t) - p(t-1);\ \gamma \text{ is discount factor, } r \text{ is external failure signal}$$

ASE EQUATIONS:
ASE output (control action fed to the polecart system):

$$y(t) = g[u_i(t)d_i(t) + noise(t)];\quad g[w] = \begin{cases} +1, & \text{if } w \geq 0 \quad \text{(right control force)} \\ -1, & \text{if } w < 0 \quad \text{(left control force)} \end{cases}$$

ASE weights:

$$u_i(t+1) = u_i(t) + \alpha\, \hat{r}(t)e_i(t);\ \alpha \text{ is learning rate}$$

ASE eligibility traces:

$$e_i(t+1) = \delta e_i(t) + (1-\delta)y(t)d_i(t);\ \delta \text{ is trace decay rate constant}$$

b

occurs, the ACE weights of the boxes that the system visited will be updated as per the equation in Table 2(b). It is clear that, since the external reinforcement signal (r) is -1, the ACE weights will always be negative. A strong negative value for ACE weight (v_i) indicates that after visiting this state, the pole cart system often failed. Whereas a value close to 0 for ACE weight, indicates that the state is associated with prolonged balancing episodes. Barto *et al.* [1] termed the weights as reflecting prediction of failure. Hence they call $p(t)$, the *prediction* signal. Alternatively, it can also be seen as the *evaluation* of a state and the resulting set of values over the state space, as a *value map*

(for a general discussion of the idea of value map in neuro-resistive grid, see [2]). According to Barto *et al.* [1], the failure signal leads to punishment of all the recent control actions of ASE that were preceding the failure and results in increasing the prediction of failure in all the recent boxes in ACE. To keep track of the visitations, a trace variable is turned on in each box whenever the system enters that box. The trace in ASE keeps track of both the nature of the action (right/left), and the length of time since that action took place. The trace in the ACE, however, does not have a sign component. ASE traces are reset on every trial but ACE traces are reset only at the beginning of a run. The effect of resetting the ACE traces every trial on the learning speed is discussed in Sect. 5. The learning of the ASE weights is accomplished by an internal reinforcement signal provided by the adaptive critic element (ACE). ACE computes the internal reinforcement signal ($\hat{r}$) by comparing the *value* of the current state and that of the previous state. Using this internal reinforcement signal, evaluations of all the visited states (i.e. the weights between decoder boxes and ACE) and actions performed (i.e. the weights between decoder boxes and ASE) are adjusted in proportion to the recency information given by the ASE and ACE traces, e_i and $\bar{a}_i$, respectively. As shown in the equations for weights in Table 2(b), the updated weights are used for calculating prediction, $p(t)$ and control action y in the next time step and are updated continuously throughout the learning period. When the learning is complete an ASE weight reflects appropriate control action for that box such that the pole remains balanced for long time periods.

The internal reinforcement signal ($\hat{r}$) is positive if the system moves from an 'unsafe' box to a 'safe box' and is negative if it is the other way round. With this signal the controller can modify both its actions and its state-evaluations continuously, and does not have to wait until the actual failure to occur before any modification can take place. Another notable feature in the evaluations made by ACE is that every move by the system will have consequences on the predictions of all the previously visited boxes in that run with the help of traces in ACE. For example, if the system moves from a 'safe' to an 'unsafe' box, all the 'live' (whose traces are turned on) ACE boxes are punished, that is, the prediction of failure is increased as a result of this move. Thus traces in ACE enable a form of generalisation across boxes. It needs to be emphasised here that each box (state) must have been visited at least once before its evaluation can be assigned. This sole fact implies a lengthy training procedure for the ASE/ACE system.

In the neuro-resistive grid approach, generalisation is achieved through lateral connections to the neighbouring states in the grid. Thus the structure of the grid itself makes the propagation of the value-information across boxes without resorting to traces in ACE. This mechanism reduces the number of computations dramatically as will be demonstrated in Sect. 5.

3. Neuro-Resistive Grid (NRG) Approach

3.1. Laplacian Methods

In this section, Laplacian methods which are the precursors to the neuro-resistive grid approach, are discussed. Laplacian function methods were introduced for robot path-planning problems (see Ref. [2] for review). Connolly *et al.* [9] proposed the use of harmonic solutions of Laplace's equation as the path lines for a robot moving from a start point to a goal point. They considered obstacles as current sources and the goal as a sink (potential fixed at zero). These conditions amount to defining *Dirichlet* boundary conditions for solving Laplace's equation. Using this solution a potential field distribution can be calculated. Given a starting point the path to the goal can be easily constructed by following current lines, which amounts to performing a steepest descent on this potential field (that is, finding a succession of points with lower potentials that lead to the lowest potential in the domain which happens to be at the goal point). It has been demonstrated by Connolly *et al.* [9] that the path constructed by using gradient descent guarantees a path to the goal without encountering local minima and successfully avoiding any obstacles. If the positions of the goals and obstacles are fixed, then the potential field has to be calculated just once and can be used for any starting point.

Tarassenko and Blake [10] modelled obstacles as non-conducting solids in a conducting medium, the starting point as a current source and the goal as an equal and opposite current sink. These conditions amount to specifying *Neumann* boundary conditions for solving Laplace's equation. As the normal derivative is fixed at the boundaries (*Neumann* conditions) of the domain, the range of values of the potential gradients will be bounded and the gradients will not decay with distance as is the case with *Dirichlet* conditions. Once solutions of Laplace's equation are found under these boundary

conditions, a potential field can be computed. A gradient descent on this potential field determines the path from starting point to the goal. Tarassenko and Blake [10] asserted that a resistive grid implementation of this method overcomes the computational problems inherent in this method.

3.2. Neural Network Implementation of Laplacian Methods

Bugmann *et al.* [2] proposed a neural implementation of the resistive grid method, the Neuro-Resistive Grid (NRG) method. The domain is discretised and mapped into the nodes of a grid of neurons connected by weights. The information of the positions of the starting point, goal and obstacles are kept in a 'memory' layer which has the same number of nodes as the neuron-grid. The neuron-grid and the memory layer are connected in a one-to-one fashion. Thus, the activation of some of the nodes in the neuron-grid can be held fixed from the memory layer. Thus each neuron i calculates its activation or potential z_i as follows:

$$z_i = f\left(\sum_{j=1}^{n} W_{ij} z_j + I_i\right)$$

where W_{ij} is the weight of the input from neuron j to neuron i, I_i is the input from the memory and $f(\cdot)$ is the activation transfer function, which is a linear saturating function such as:

$$f(\zeta) = \begin{cases} 0 & \text{if } \zeta < 0 \\ \zeta & \text{if } 0 < \zeta < 1 \\ 1 & \text{if } \zeta > 1 \end{cases}$$

By using $W_{ij} = 1/n$, where n is the number of neighbours ($= 2m$ where m is the dimension of a square grid), the neurons set their potentials to the average potential of all the neighbours. Bugmann *et al.* [2] have shown that this formulation leads to a Poisson equation and the solution of this Poisson equation determines the activation (potential) distribution of the neurons in the NRG. The shape of the potential distribution in NRG depends on the encoding scheme of the forbidden states (obstacles), whether the corresponding nodes are current sinks (Dirichlet condition) or are disconnected from the grid (Neumann condition) [2].

4. Application of NRG to the Pole-Balancing Problem

In the work reported here the ACE is replaced by an NRG. As described in Sect. 2, ACE develops a value map of the system states. The rationale for replacement is that by propagating the failure information across the grid (especially to the immediate neighbours), in the future if the system enters this neighbourhood, the NRG will send a failure prediction signal that enables the ASE to adjust the recent control actions performed. This in turn prevents the system from performing a similar sequence of actions in the future. Thus, even if some states are not visited in the previous trials, the NRG system deems them as safe if they are away from 'bad' boxes and unsafe if they are near a 'bad' box. Whereas in ACE, if a particular box is not visited so far in a run, there is no prediction available for that box. Thus in the NRG system it is possible to build up a value surface relatively quickly.

As shown in Fig. 1, a four-dimensional resistive grid is constructed for the four variables, position (x), angle (θ), velocity ($\dot{x}$) and angular velocity ($\dot{\theta}$). In Fig. 1, the neighbourhood for one cell is shown schematically. The control system set-up that includes the NRG, is shown in Fig. 2. Apart from the external input (I_i) from the memory layer, a

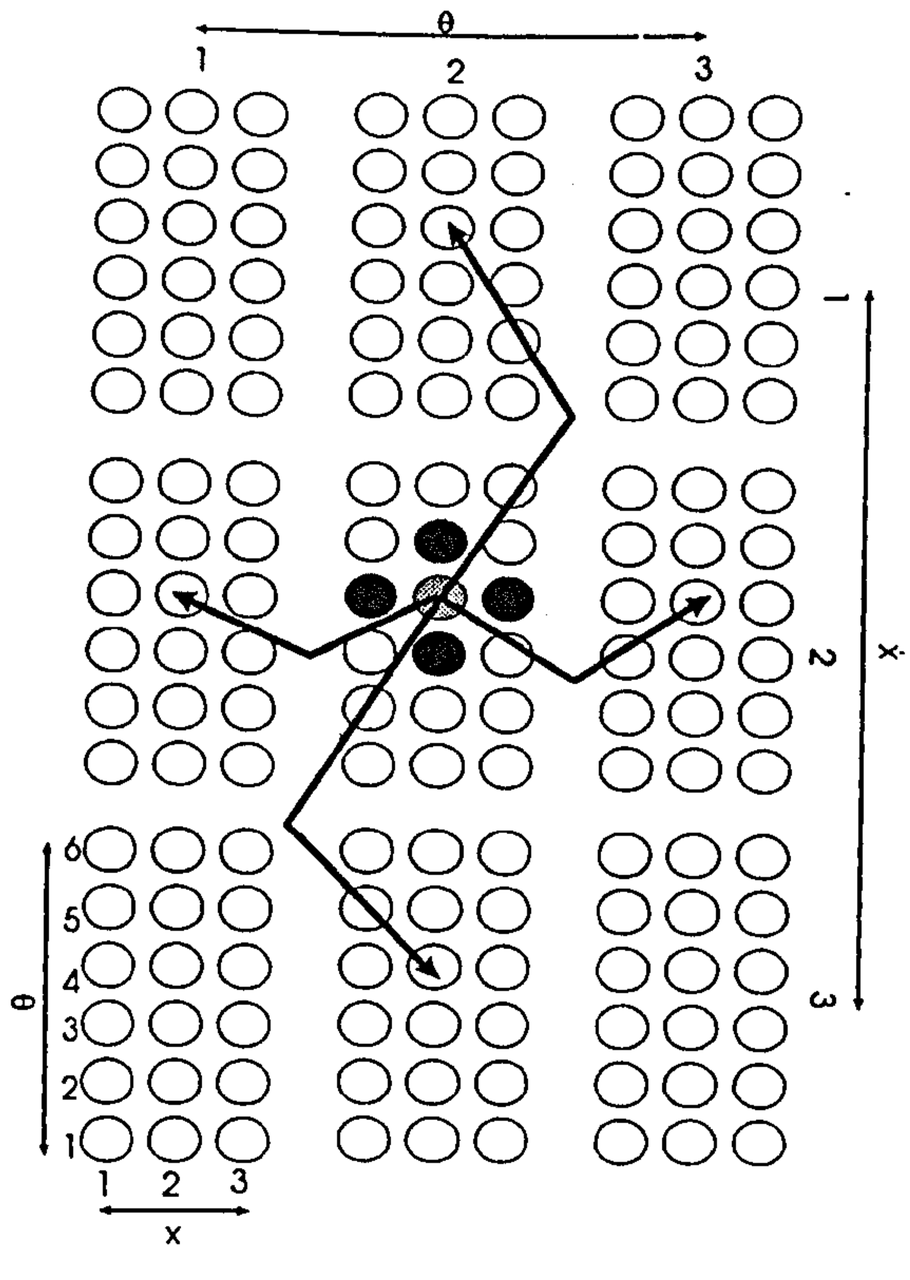
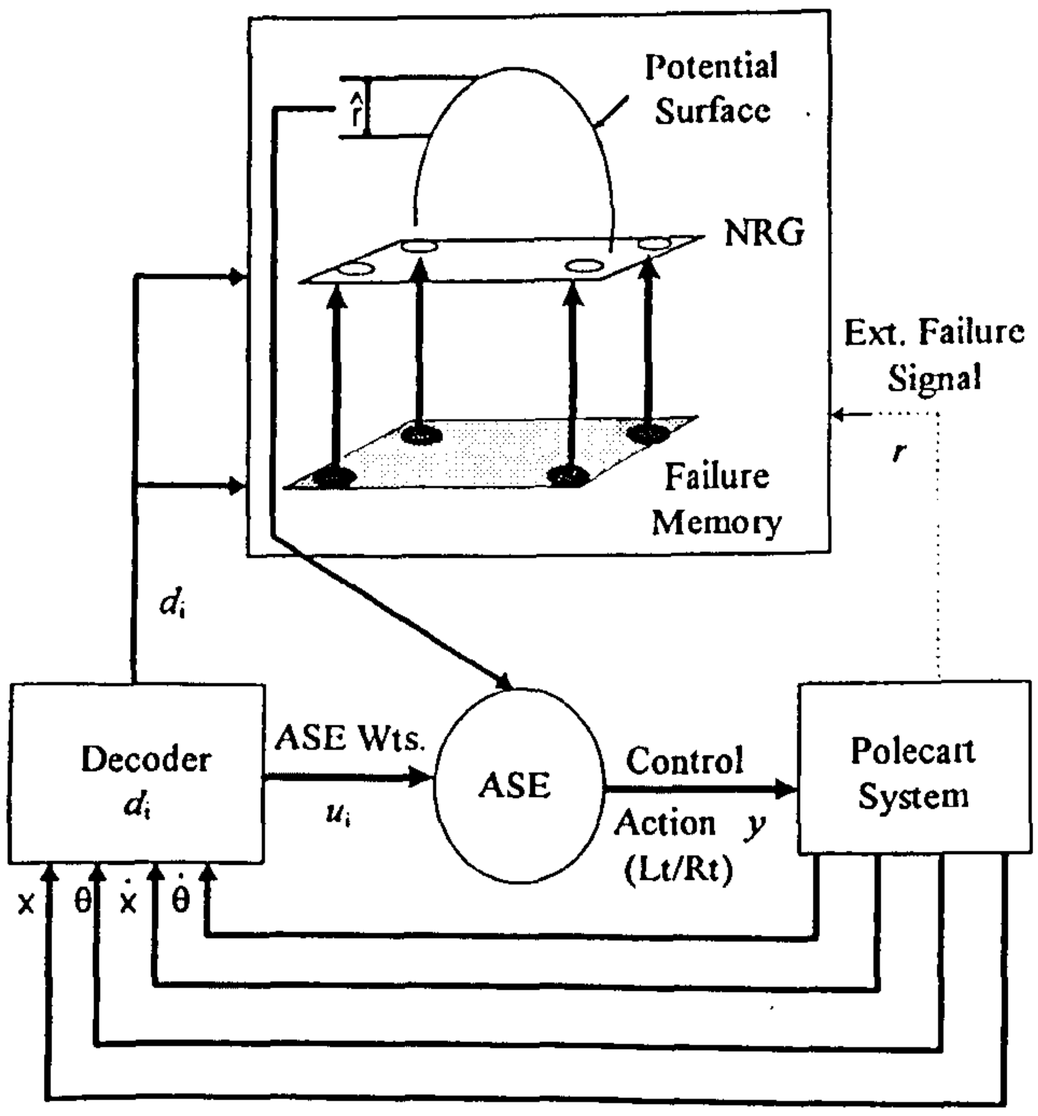


Fig. 1. 3×3 grid showing the arrangement of the 162 boxes that constitute the pole cart system state space. The convention for the variables is shown on the sides of the grid and the numbers represent the indices of the quantised regions for each variable. An example of connectivity for one node is illustrated. The node here has eight neighbours. The nodes on the edges have less number of neighbours. The connection strength is fixed so that the edge affects are balanced.



(a)

$\hat{r}$ = η(New Value − Old Value)	
$\hat{r}$ Value	State Transition
> 0	'Bad' to 'Good'
< 0	'Good' to 'Bad'
0	'Bad' to 'Bad' 'Good' to 'Good'

(b)

Fig. 2. (a) The NRG/ASE system configuration showing the replacement of the Adaptive Critic Element (ACE) by the Neuro-Resistive Grid (NRG) (compare from the adaptive critic system in Table 2(a)). The internal reinforcement signal is fed to the Adaptive Search Element (ASE) which in turn generates a control action; (b) table indicates the interpretation of the internal reinforcement signal. The reinforcement signal is set equal to the difference between the *values* of the new and the old boxes that the system entered and η is a constant equal to 0.95.

small bias voltage (0.01) is fed to all the nodes in the grid. Incorporation of the bias voltage is a novel modification of the original NRG technique [2] to adapt it to pole balancing application, as discussed below.

In the application of NRG to path planning, there is a goal state, which is a node with the highest potential that the system is asked to attain. However, the formulation of the pole balancing problem does not define a goal state for the pole. It is only specified that falling should be avoided. Therefore, we have assigned an initial goal value to each state of the pole-cart system by providing each node in the grid with a small current source produced by the bias voltage in the memory layer. States where the pole falls during training are transformed into current sinks. This progressively leads to a potential distribution with highest values for states where the pole is close to a vertical and centred position, as in the ACE evaluation function.

The weights between neighbours are set to $W_{ij} = (1 - \text{bias})/n$ where n is the number of neighbours and bias is set to 0.01. This bias term in the weights prevents saturation of node voltages. At the beginning of every run the grid is updated (i.e. every node activation is replaced by the average activation over its neighbourhood) until the potential field stabilises. In our simulations we observed that 30 cycles of updating achieves a stable potential field. The grid is not updated again until a failure occurs. On failure, the node corresponding to the failure is set to a potential of −1 in the memory (shown as 'failure memory' in Fig. 2) layer and the grid is cycled for 30 times. A run is terminated if the system reaches a cumulative time count of 500 000 time steps or if there are 100 failures (trials), whichever happens first (the same scheme was used in Barto *et al.* [1]). This amounts to a total simulation time of balancing approximately equal to 2.8 hours. The actual simulation, performed on Cortex-Pro neural network simulation package running on a PC-486 DX2-50, took approximately eight hours for the NRG algorithm. It is interesting to note that the adaptive critic algorithm took on an average about 18 hours to complete a run.

At each time step, internal reinforcement ($\hat{r}$) is computed by taking the difference between activation (in the resistive grid) of the present and previous nodes (boxes) that the system entered. This signal in turn adjusts all those ASE weights whose eligibility traces are active. In the adaptive critic approach the ACE predictions are also adjusted every time step, whereas the potential field in NRG is updated again only after a failure. Thus the NRG approach is computationally more economical than the adaptive critic approach. The number of updates of NRG is a constant proportional to the number of failures whereas the number of updates in ACE is proportional to the number of time steps of balance (which can grow exponentially once the system stays balanced). When the system fails in the NRG approach, the node that the system entered immediately before the failure is noted. This information is made available to the failure memory layer, where this node is clamped to −1 for the rest of the simulation. During the grid cycling period, this

information gets propagated to the neighbouring nodes. Hence, a potential surface reflecting the relative goodness of a state (value map) evolves in the NRG. As in the system of Barto *et al.* [1], the NRG system also produces a reinforcement signal with the same interpretation. If the system transition is from a 'good' to a 'bad' state the reinforcement signal is negative indicating a punishment signal to all the recent actions of ASE. Whereas, if the transition is from a 'bad' to a 'good' state, the reinforcement signal is positive indicating a reward signal to all the recent actions of the ASE. The NRG system sends a neutral signal to other transitions (see Fig. 2(b)).

The quantisation of the state space is prefixed in both the adaptive critic and the neuro-resistive grid algorithms. Performance improvements have been reported with adaptive quantisation schemes (for example, see Ref. [4]). In these schemes the boundaries of regions that make up system states are made elastic so that they contract or expand based on the behaviour of the system. At the end of training and stabilisation, important regions will have been quantised with a finer resolution and others with a coarse resolution. Thus, performance of the adaptive critic algorithm can be enhanced by using an optimal partitioning scheme of the state space. Since the basic mechanism of the NRG algorithm, namely the propagation of information to neighbouring nodes to enable faster generalisation, does not depend on the particular quantisation scheme used, all the advantages of using an optimal quantisation scheme discussed above are transferable to the neuro-resistive grid algorithm. Hence, the relative improvements over the adaptive critic algorithm reported in the next section still hold true with a different partitioning scheme.

In the next section simulation results are presented. We are aware of many other approaches to pole balancing (reviewed in Ref. [3]). However, since the efforts in this work are aimed at introducing new ways of evaluating states as described above, all the comparisons of performance in the next section, will be made with reference to the system of Barto *et al.* [1].

5. Results and Discussion

Both the systems are run with the benchmark parameters (note especially the increased angle of failure) and the system equations as advised in Ref. [3] (indicated in Table 1). Geva and Sitte [3] felt that the previous failure angle of 12° is too restrictive, and advised for a 90° failure angle for the benchmark experiment. Both the systems are simulated for 10 runs. All the trials start from the zero initial conditions on the variables. In all of the runs, the system learnt successfully before completing 100 trials (failures). Hence all the runs were terminated before 100 trials. To avoid having runs of different lengths, all the remaining trials in a run are assigned a value equal to the higher of the two immediately preceding times of balance [1]. For each run a data point is plotted by averaging the number of time steps to failure in five successive trials (see Figs 3(a), (b)). Then these results are averaged over all the 10 runs (see Fig. 4).

Figures 3(a) and (b) depict the results for the adaptive critic and NRG, respectively, for each of the 10 runs. Figure 4 shows the rate of learning for the two systems. The graphs clearly show that NRG is able to learn a successful controller quickly. For example, by about 15 trials an average adaptive critic controller is capable of balancing for about 120 000 time steps whereas an average NRG controller can balance for about 250 000 time steps. Also, the rate of learning is faster for NRG compared to that of the adaptive critic. NRG starts learning very

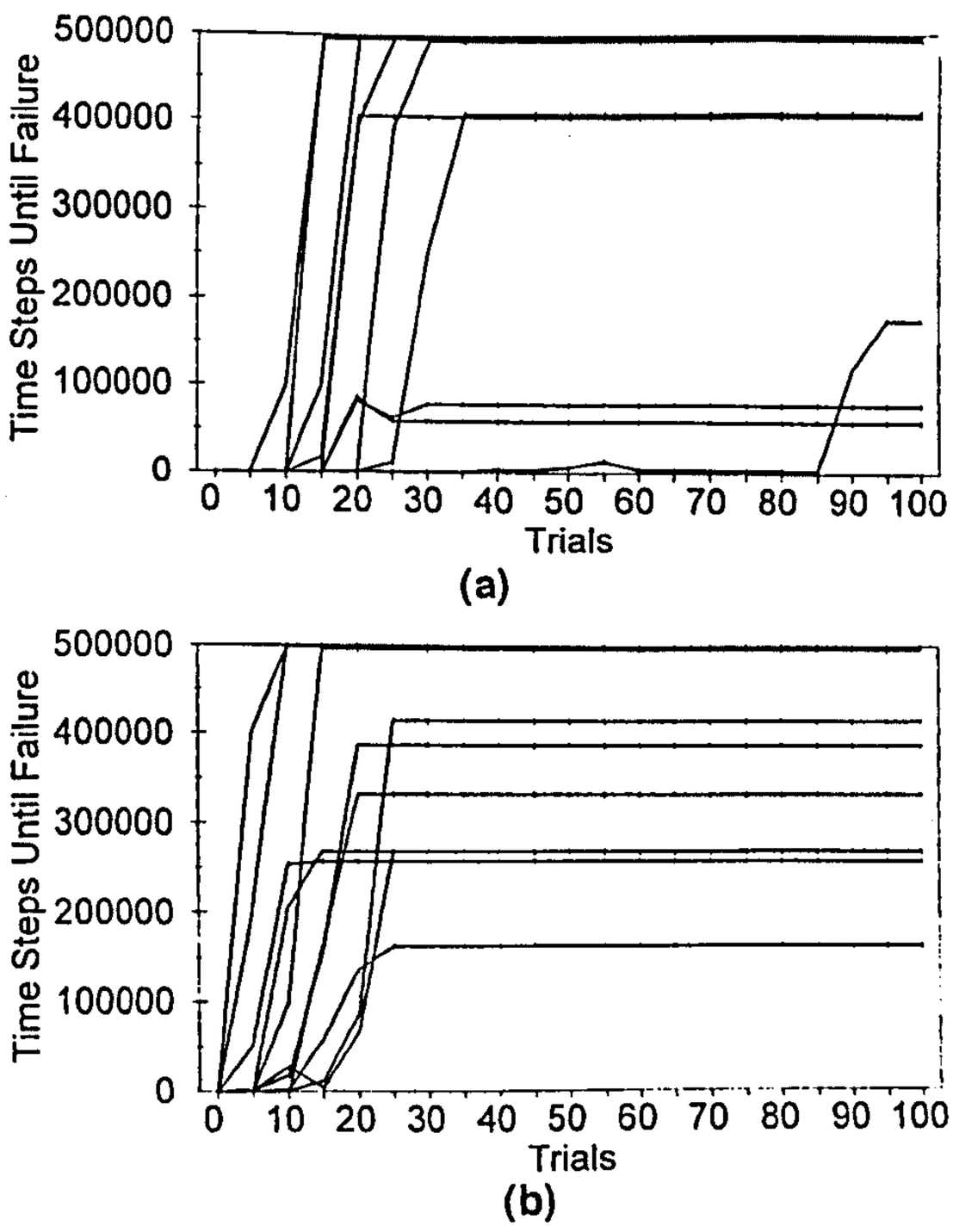


Fig. 3. Results of the training phase of the adaptive critic and the NRG algorithms. Time steps to failure for each of the 10 runs with the (a) adaptive critic approach, and (b) the Neuro-Resistive Grid (NRG) method are shown. Each data point is obtained by taking an average of time steps of balance over five trials.

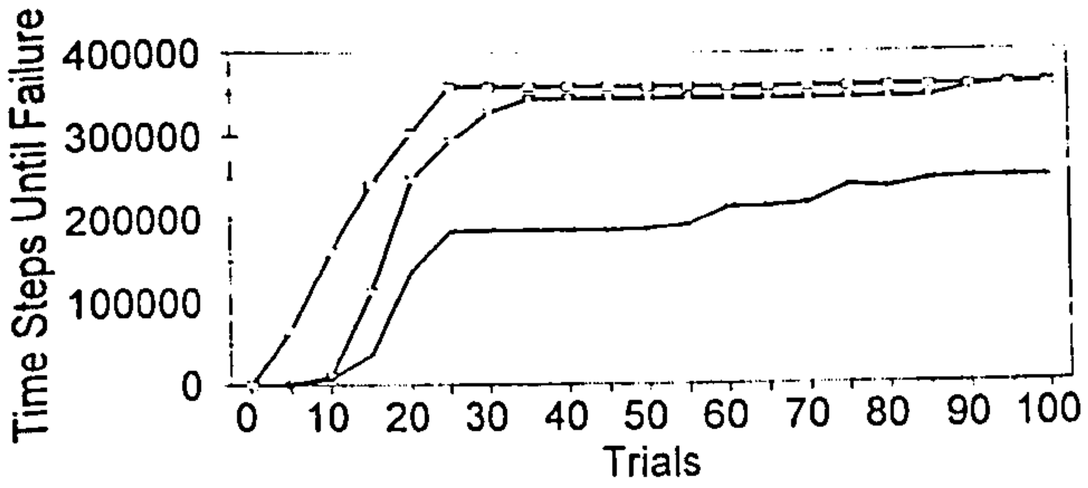


Fig. 4. Learning curves comparing the speed of learning in the adaptive critic method versus the NRG method. Each data point is obtained by taking average over the 10 runs shown in Fig. 3. The graphs clearly show that NRG is able to learn a successful controller quickly. Two graphs are shown for the adaptive critic algorithm, one with the ACE traces reset at the end of every run, and the other with these traces reset at the end of every trial. The former learns faster than the latter and both of these are slower than the NRG algorithm. ▽: adaptive critic (reset end of run); ×: adoptive critic (reset end of trial); □: NRG.

rapidly due to the availability of the failure information through lateral connectivity in the grid. In the original Barto *et al.* [1] algorithm, the temporal traces in ACE are reset only at the end of each run. To investigate the role of the traces in the ACE, we ran the simulations of the adaptive critic algorithm wherein the ACE traces are reset at the end of every trial. From Fig. 4 it is clear that the learning speed is the lowest if the ACE traces are reset at the end of every trial. By waiting to reset the traces only at the end of a run, the states that were visited in the previous trials will remain active in the subsequent trials. It appears that this facility enables a continuous adjustment of the evaluations, and thus leads to faster learning in the adaptive critic algorithm, as demonstrated by the learning curves in Fig. 4. In the NRG algorithm, this process is improved further in that the states do not have to be necessarily visited before their evaluation is ascertained. Due to the lateral connectivity, every state will have an evaluation based on the neighbourhood. We suspect that this lateral diffusion of information is critical to NRG's success.

To characterise the performance of both the controllers, two benchmark tests [3] are conducted. The first test determines the capability of the controller to balance the pole from a different starting position other than that of the training period and also characterises the amount of deviation of the angle of the pole from the vertical and that of the cart position from the centre of the railing. The new initial conditions are: $x = -1$, $\theta = -0.1$, $\dot{x} = -1$, $\dot{\theta} = -0.2$. The values of the cart position and the pole angle over the first 1000 time steps are shown in Figs 5(a) and 6(a) and the values for the rest of the 50000 time steps are shown in Figs 5(b)–(c) for

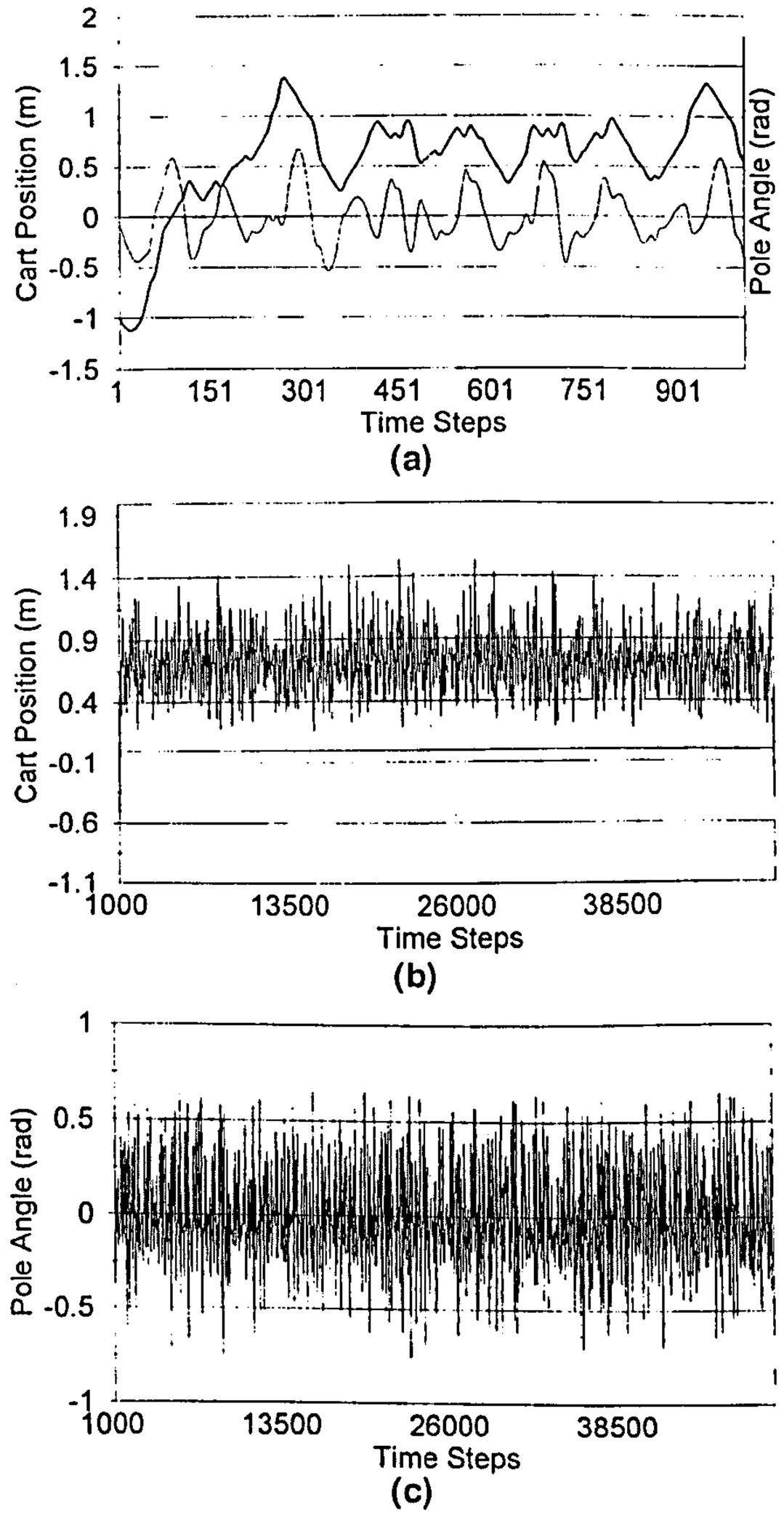


Fig. 5. Results of the first benchmark test on the adaptive critic-controller. (a) Graph shows the cart position (—) and the pole angle (—) over the first 1000 time steps. The Root Mean Squared (RMS) value and the Standard Deviation (SD) for the cart position are 0.76 m and 0.49, respectively. The RMS and SD values for the pole angle are 0.27 radians and 0.27, respectively; (b) graph shows the deviation in the cart position for the remaining time steps till 50000. The RMS and SD values for the cart position are 0.76 m and 0.49, respectively; (c) graph shows the deviation in the pole angle for the remaining time steps till 50000. The RMS and SD values for the pole angle are 0.28 radians and 0.28, respectively.

the adaptive critic system and in Figs 6(b)–(c) for the NRG system. The controller trained by NRG has smaller deviation in both the position and the angle throughout the test period. However the controller trained by the adaptive critic method has larger deviation in these values. This indicates that the AC-controller allows big oscillations both in

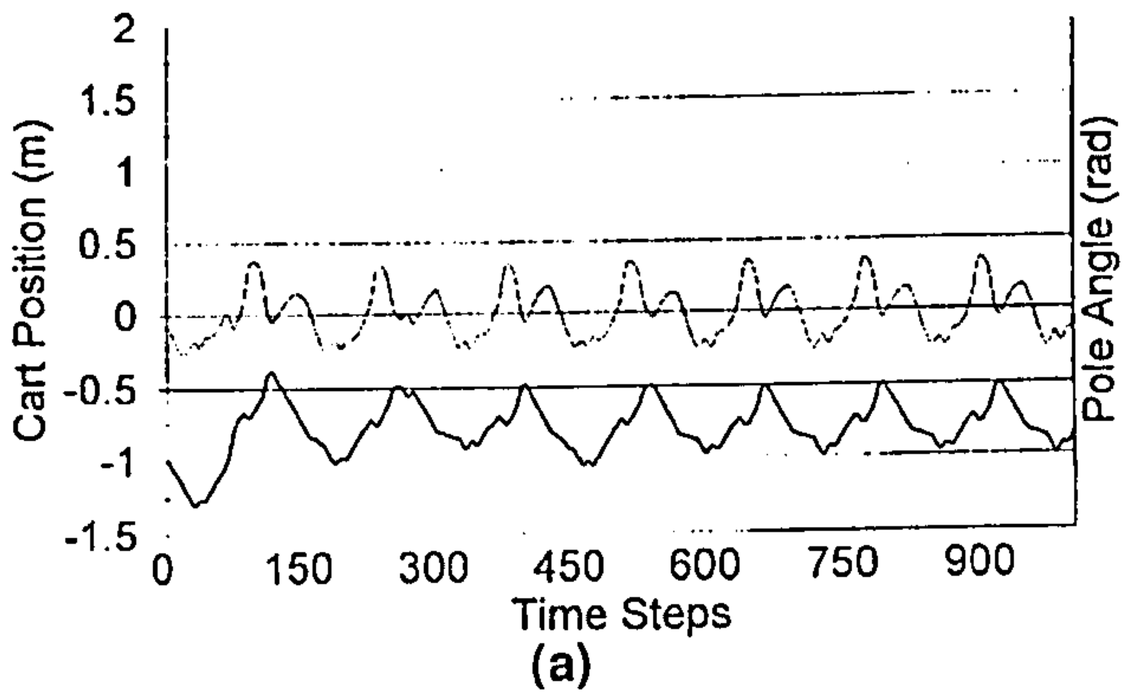
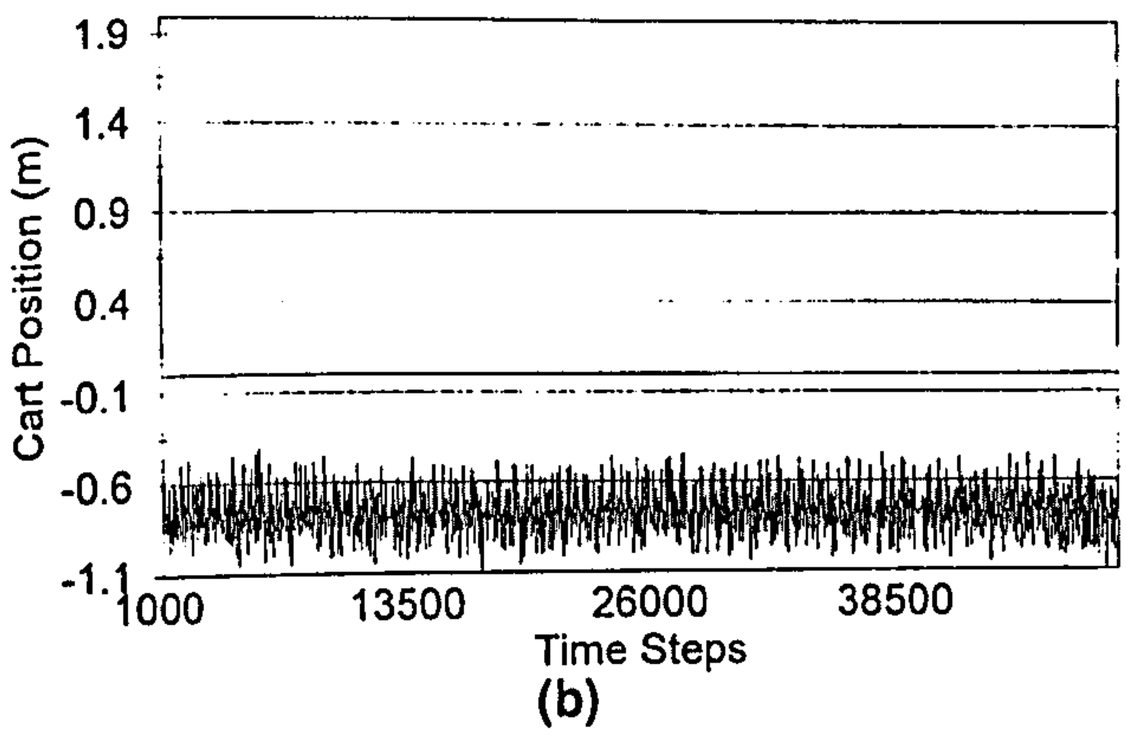
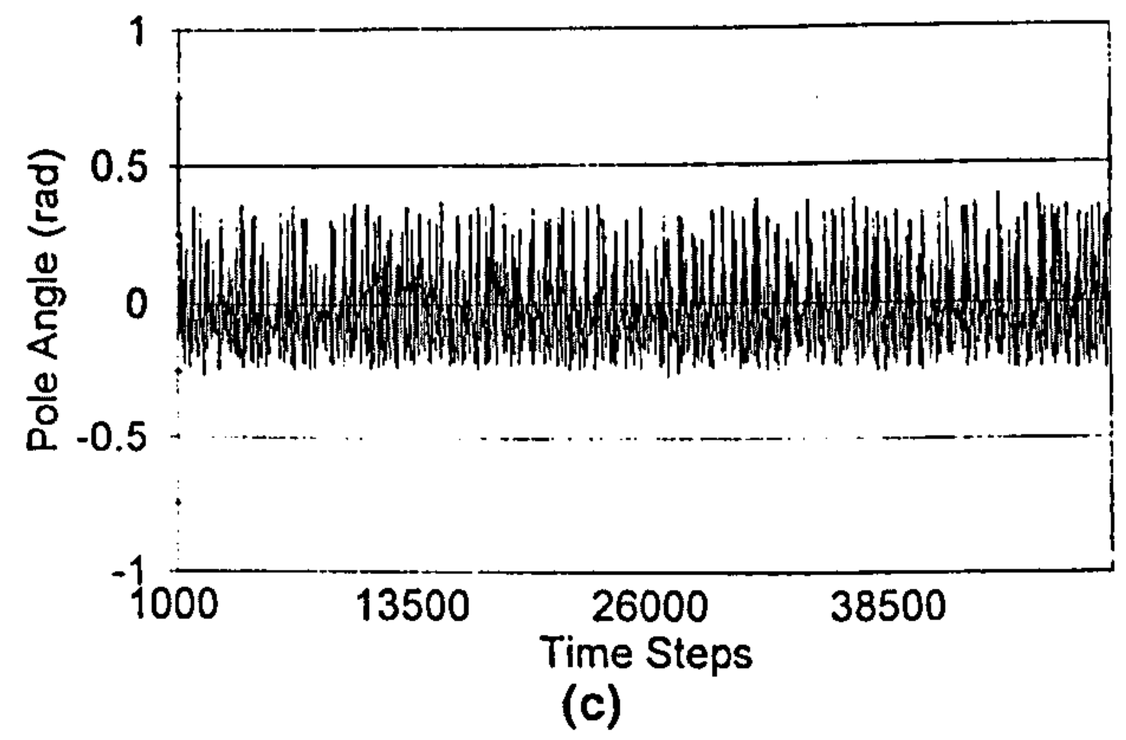


Fig. 6. Results of the first benchmark test on the NRG-controller. (a) Graph shows the cart position (—) and the pole angle (–) over the first 1000 time steps. The Root Mean Squared (RMS) value and the Standard Deviation (SD) for the cart position are 0.81 m and 0.18, respectively. The RMS and SD values for the pole angle are 0.17 radians and 0.17, respectively.



(b) graph shows the deviation in the cart position for the remaining time steps till 50000. The RMS and SD values for the cart position are 0.8 m and 0.16, respectively.



(c) graph shows the deviation in the pole angle for the remaining time steps till 50000. The RMS and SD values for the pole angle are 0.17 radians and 0.17, respectively.

position and angle whereas the NRG-controller does not. Both the systems have residual oscillations because of the bang-bang nature of the control force.

The second test characterises the dynamic range of the controller. The pole cart system is released from various initial angles and angular velocities, with the cart at the centre of the track. If the system remains balanced for 15000 time steps, a data point corresponding to the pole angle and the angular velocity is plotted on the graph. Thus a plot of all such points reflects the range of initial conditions from which the pole can be balanced successfully for long time. This range is termed as the dynamic-range and it reflects a form of generalisation over the set of initial conditions. We have noticed that not all the trained controllers were successful on the first benchmark test. So we chose the successful controller on the first test that possesses the largest dynamic range as the best controller for each of the adaptive critic and the NRG systems, for comparison in Fig. 7. It is evident from the graph that the NRG-controller has less dynamic range compared to the adaptive critic-controller.

In summary, the first test demonstrates that the NRG-controller allows less deviation in position and angle. The second test reveals that the range of initial conditions over which an NRG-controller can successfully balance is smaller than that of the adaptive critic-controller. Although the NRG system learns a controller faster than the adaptive critic system, the dynamic range is more limited. A comparative summary of performance characteristics of the two algorithms is compiled in Table 3.

These results suggest a possible trade-off between the learning speed (computational expense) versus

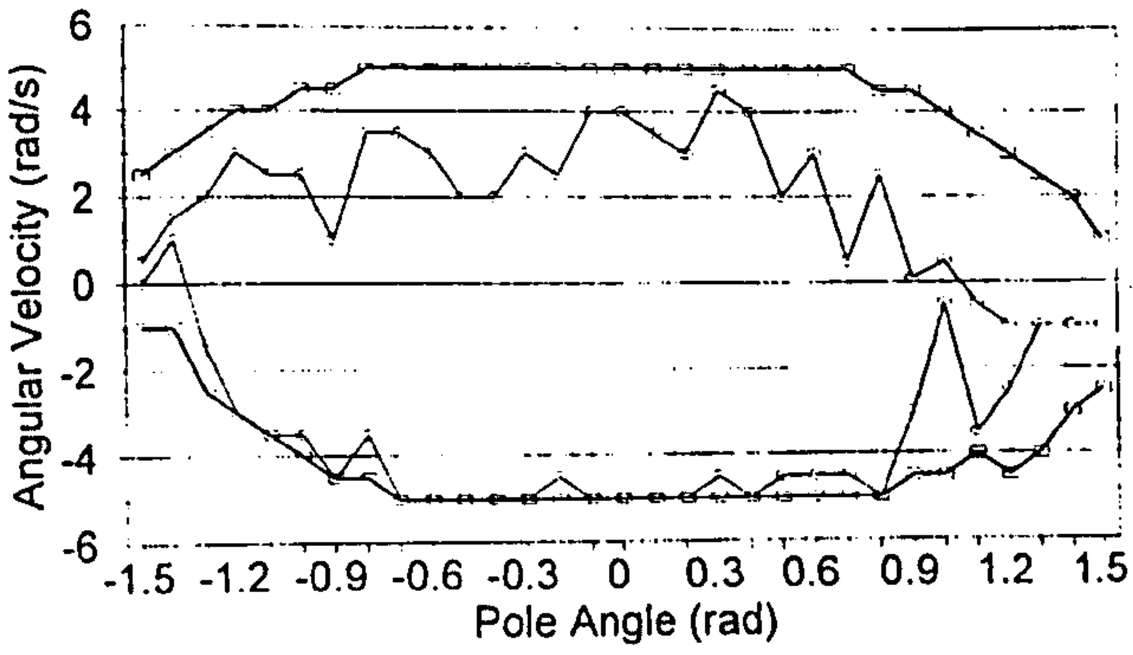


Fig. 7. Results of the second benchmark test on both the controllers. The pole cart system is released from the centre of the track each time and the initial angle and angular velocity of the pole are varied. For each angle of the pole, the minimum and maximum angular velocities at which the controller can successfully balance for 15000 time steps are plotted. Thus the dynamic range for the adaptive critic (□) versus NRG (○) controllers is characterised. The NRG controller has lower dynamic range than the adaptive critic controller (see text for discussion).

Table 3. Comparison of the NRG and the adaptive critic algorithms for control.

Property	NRG	AC
Learning speed	Fast	Slow
Number of computations	Less	More
Quality of control	Less angular/position deviation	More angular/position deviation
Dynamic range	Medium	Large

the dynamic range (generalisation over the set of initial conditions). We observed in our simulations of NRG that some of the ASE weights are either not modified at all (because the system never visited these states) or they are very small, of value comparable to the noise term, the latter being due to successive small positive and negative weight changes which cancel each other. These small or unset weights are a probable cause of lack of dynamic range which we discuss below as due to a 'lack of experience' and due to 'rigidity of value map'. Another possible reason for limited dynamic range is a non-optimal encoding of sequences of visited states in the weights. This is discussed below as the 'relearning problem'.

5.1. Lack of Experience

The NRG converges very rapidly to an almost correct evaluation of states and thus provides correct feedback very early to the ASE. Thereby, for each visited state the ASE learns a correct action and as soon as a sequence of actions leads to a cyclic pattern of visited states, learning stops as no new states are visited. Thus, although ASE is being successful quickly, it has less knowledge of control actions over the state space.

While NRG learns faster than ACE due to lateral propagation of values, the ASE still relies on the technique used by Barto *et al.* [1], wherein learning takes place only for the states that are actually visited. This suggests the need for matching convergence speeds of ACE/NRG and ASE systems. We have observed in our simulations that the weights of ASE have similar values in neighbouring regions of the state space. Hence, it may be possible to increase the learning speed in ASE also by virtue of lateral propagation of information. Thus, achieving generalisation over actions may probably call for a new way of representing actions in the ASE system.

In this context, we wish to alert the reader that incorporation of prior knowledge must be considered with caution. Prior knowledge can be incorporated in ACE/NRG by prefixing the evaluations of the edge states where we know failures occur. However, by doing this, ACE/NRG learning may be further accelerated and as a result the dynamic range may be reduced even more because by learning faster the system may not have had opportunity to visit many states.

5.2. Rigidity of Value Map

In the NRG the values of states are not modified within a trial, i.e. until failure occurs. On the one hand, this reduces computational expense, but on the other hand, it also leads to a rigid value map during a trial which may contribute to cancellation of positive and negative weight changes. In contrast, in the ACE valuations of states are modified at each step so that returning to a previous state will not, in general, result in an internal reinforcement signal ($\hat{r}$) of the same value but opposite sign leading to similar cancellations as in the NRG. It remains to be verified how important this effect is.

5.3. Relearning Problem (of Evaluations for Failed States)

In the NRG method, a failed state remains clamped as failed for the rest of the run. It must be noted that the quantised states used here cover a large portion of the system state space. Depending on how a state is entered, a given action may lead to a successful sequence of movements, while another action may lead to a failure. Thereby a too early definitive evaluation of a state may prevent a number of potentially successful sequences of states to become part of the dynamics encoded in the ASE weights. In contrast, the adaptive critic approach allows a failed state to alter its value if it becomes part of a viable sequence of balancing forces in the future. This appears to be a subtle problem which deserves further analysis.

6. Conclusions and Future Directions

In summary, an NRG algorithm is presented for the pole-balancing problem and compared to the ACE

algorithm. Both the systems are simulated using the same benchmark parameters. Results from the training phase show that the NRG algorithm learns faster and with fewer updates than the adaptive critic algorithm. This is due to the propagation of failure information to the neighbouring nodes in the NRG. Benchmark tests are conducted to compare the performance of the controllers trained with both the algorithms. Firstly, it is observed that the controller discovered by NRG keeps the angle and position of the pole in a narrow range, whereas that learnt by the adaptive critic method allows more variation in both these variables. Secondly, the dynamic range (range of initial conditions for which the controller can balance the pole for long periods of time) of the NRG-controller is relatively less than that of the adaptive critic-controller.

Directions for future investigation have been outlined that lead toward a better understanding of the reasons for the differences and improvements of the NRG controller. There are two potential causes for the reduced dynamic range of the NRG controller. Firstly, it may be that the NRG-controller lacks sufficient experience as the system learns successfully before the occurrence of many failures. This may be compensated for by training the system with a wide range of initial conditions rather than training from a fixed initial position. Alternatively, it may be possible to reduce the mismatch between convergence speeds of the evaluations in the NRG and the action policy in the ASE by using a lateral propagation scheme in the ASE as well. Secondly, it may be due to the absence of flexibility in the evaluation of the states. In the current system when once failure occurs in a state, that state is fixed as 'failed' throughout the run. One way to improve on this is to allow flexibility in the memory layer so that a failed state can become 'good' in future trials if there is sufficient evidence from the behaviour of the system.

In recent times, temporal difference methods have been successfully applied to many interesting control problems such as the mountain car problem [12], backgammon [13], automatic aircraft landing [14], etc. Exploring the application of NRG to these problems may help gain insight into both the NRG and TD methods.

On the theory side, unified view of dynamic programming, reinforcement learning and heuristic search has been proposed [7]. Most of these advances are in the area of estimation of optimal value function. The formal link between TD learning and resistive grid method remains to be investigated.

Acknowledgements. Grant (GR/J42151) support from the Engineering and Physical Sciences Research Council (EPSRC), UK to RSB (post doctoral research fellowship) and BDC (research studentship) is gratefully acknowledged.

References

1. Barto AG, Sutton RS, Anderson CW. Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans Syst, Man & Cybern 1983; 13: 834–846
2. Bugmann G, Taylor JG, Denham MJ. Route finding by neural nets. In: Taylor JG (ed), Neural Networks, Unicom & Alfred Waller, UK, 1995, 217–231
3. Geva S, Sitte J. A Cartpole experiment benchmark for trainable controllers. IEEE Control Systems Magazine 1993; 13: 40–51
4. Rosen BE, Goodwin JM, Vidal JJ. Process control with adaptive range coding. Biol Cybern 1992; 66: 419–428
5. Sutton RS. Learning to predict by the method of temporal differences. Machine Learning 1988; 3: 9–44
6. Barto AG, Sutton RS, Watkins CJCH. Learning and sequential decision making. In: Gabriel M, Moore J. (ed.), Learning and Computational Neuroscience: Foundation of Adaptive Networks, MIT Press, Cambridge, MA, 1990, 539–602
7. Barto AG, Bradtke SJ, Singh SP. Learning to act using real-time dynamic programming. Artificial Intelligence 1995; 72: 81–138
8. Ribeiro CHC. Attentional mechanism as a strategy for generalisation in the Q-learning algorithm. In: Fogelman-Soulié F, Gallinari P. (ed.), Proc. ICANN '95, Paris, 1995; 1: 455–460
9. Connolly CI, Burns JB, Weiss R. Path planning using Laplace's equation. Proc IEEE Int Conf Robotics & Automation 1990; 2102–2106
10. Tarassenko L, Blake A. Analogue computation of collision-free paths. Proc IEEE Int Conf on Robotics & Automation, Sacramento, CA, 1991, 540–545
11. Sutton RS, Pinette B. The learning of world models by connectionist networks. Proc Seventh Ann Conf of the Cog Sci Soc, Lawrence Erlbaum, 1985, 54–64
12. Moore AW. Efficient memory-based learning for robot control. PhD thesis, University of Cambridge, 1990
13. Tesauro G. Temporal difference learning and TD-Gammon. Comm ACM 1995; 38(3): 58–68
14. Prokhorov DV, Santiago RA, Wunsch II DC. Adaptive critic designs: A case study for neurocontrol. Neural Networks 1995; 8(9): 1367–1372